ABOUT THE AUTHORS

Dr. Ajit Vijaykumar Patil completed his B.A.M.S. in 2012 and M.D. (Samhita Siddhant) from Hon. Shri. Annasaheb Dange Ayurved Medical College, Post Graduate & Research Centre, Ashta in 2015.

Dr. Pramod Prakash Kanap completed his B.A.M.S. in 2011 and M.D. (Kriya Sharir) from Hon. Shri. Annasaheb Dange Ayurved Medical College, Post Graduate & Research Centre, Ashta in 2014. After PG he joined as a Lecturer in A.D.A.M.C, Ashta in 2014 and continues till date. Also he is an associate editor of "SUBHADRA – International Journal of Ayurveda"

SALIENT FEATURES

- Strictly as per new syllabus of B.A.M.S prescribed by C.C.I.M
- Easy,Simple & Lucid Language.
- Diagrammatic & Graphical Presentation.
- Maximum statistical problems solved methodically.
- Reviews & questions of each chapter elaborated.

Chaukhamba Surbharati Prakashan
Varanasi-221001
csp_naveen@yahoo.co.in

ISBN : 978-93-85005-35-0

Research Methodology & Medical Statistics

Research Methodology & Medical Statistics

[Text Book for B.A.M.S. - 4th Prof.]

Dr. Ajit Patil
&
Dr. Pramod Kanap

THE
CHAUKHAMBA AYURVEDA STUDIES
21

Research Methodology & Medical Statistics

[Text Book for B.A.M.S. 4th Prof.]

by

DR. AJIT PATIL
M.D. (Samhita Siddhant)

&

DR. PRAMOD KANAP
M.D. (Kriya Sharir)
Assistant Professor,
Deptt. of Kriya Sharir
A.D.A.M.C, Post-Ashta
Sangli – (Maharastra)

Chaukhamba Surbharati Prakashan
Varanasi

Ashta these great persons and experts helped us for the collection of information. They give us a proper guidance about the collection of information.

While doing all these things our parents, family members and friends, they all helped us and promote us. We are very thankful to them. Also we are thankful to Dr. Shital Kanap & Dr. Vijay Dange for helping us. We express our indebtedness to all those who helped us directly and indirectly.

And Last but not the least: we are very much grateful to Choukhambha Publication and Mr. Neeraj gupta for giving us this opportunity.

While writing this book we take the guidance of Research and Statistics subject experts. If there are some mistakes then please convey it. We welcome all their comments and progressive suggestions about book. We are assuring that we consider them in next edition.

DR. AJIT V. PATIL
DR. PRAMOD
P. KANAP

CONTENTS

PART - I

PART-II

• • •

7. The concept and importance of ethics in research
8. Concept of Evidenvce Based Medicine and Scientific Writing
9. Importance of IT in data mining and important research data portalsconcerned with Ayurved and contemporary medical science (DHARA , PubMed, Ayush Research Portal, Bioinformatics Center, Research Management Informatic System etc.)

Part – B Medical Statistics

1. Definition, scope and importance of the Medical statistics
2. Common statistical terms and notations
   a. Population
   b. Sample
   c. Data
   d. Variable
   e. Normal distribution
3. Collection and Presentation of data
   a. Tabular
   b. Graphical
   c. Diagrammatical
4. Measures of location
   a. Average
   b. Percentile
5. Measures of Central Tendency
   a. Arithmetic mean
   b. Median
   c. Mode
6. Variability and its measurement
   a. Range
   b. Standard deviation
   c. Standard error
7. Introduction to probability and test of significance
8. Parametric and non-parametric tests
9. Introduction to commonly used statistical soft-wares.

• • •

Chapter 1

HISTORICAL BACKGROUND OF RESEARCH

Ayurveda is very eternal science. Propagation of Ayurveda on the earth starts from god Bramha to Daksha prajapati to Ashwini kumara to Indra.

In human being it went on progression and propagating by the tireless efforts of Acharya Punarvasu Atreya, Devodas Kashiraj (Dhanvantari), Bhardwaj Rushi, Charaka, Sushruta, Vagbhata, Nimi, Nagarjuna etc.

In progression of ayurveda science, ancient period had so many innovations, new ideas, new medicines, methods and many other things added by scholars in ayurveda by theirs studies and research.

1. In acharya charak period mainly ayurvedic medicine branches were highly developed.
2. In acharya sushruta period various types of operatives were done.
3. In acharya Nagarjuna period various rasa-kalpas were prepared and discovered.

Already all these medicine procedures, everything has been proved effective by their own experiments, experience, observation and results on their patients. But now in globalization era we are again proving this by latest research methodology techniques.

(1) Brief Historical Background of Research in Ayurveda

Researches in ancient ayurveda is :

A. Miracles of Ashwini kumar

1. Acharya ashwinikumar has discovered miraculous surgeries like transplanting testis of goat to lord Indra.
2. Vishala wife of king khela when got injured, she was operated by prostatic implant of iron cast.
3. Treated effectively Rajyakshma (tuberculosis) of somraja very effectively.

4. With help of rasayan chikitsa he gave very younger and useful life to Chyavan rushi. (Anti-aging therapy)

B. Charak

The various ayurvedic medicines used by acharya charaka and describing chikitsa sthana are experience oriented research of that time.

C. Sushruta

1. Acharya sushruta has described macro and micro structures of various organ of the body; for garbhsharir he himself dissected many cadavers and hence study thoroughly.
2. Acharya sushruta described various surgeries meticulously e.g. asthibhagna (Fracture of bones), linganash (cataract), nasa sandhan (rhinoplasty) at that time & now in modern days also many procedures of sushruta are still used so this ayurvedic scholar is called as "Father of Surgeries."

D. Acharya Nagarjuna

1. In revolution, ayurvedic therapy is done by using Ras, Maharasa, Ratna, Upratna, Dhatu, upadhatu, Vishdravya and detoxification (shodhan) etc. is palatable for human being.
2. Even toxic material like parada, gandhka he detoxified it with various sanskaras and made it use in various diseases.

E. Bhaishjya Kalpana

Now days various forms of medicines like tablets, capsule, syrup, decoction, ointment, nasal drops, etc. it all described in ayurvedic Bhaishjya kalpana like swarasa, kwatha, avaleha, churna, malahar, vati, guti etc.

Brief Historical Background of research in contemporary science

Ayurveda and other health sciences like Allopathy, Homeopathy and Naturopathy etc. are also developed step by step by various experiments and observations.

A. Research in Allopathy

1. Function and structure of various systems like function of various hormones of various gland, blood circulation tract, digestion processes, respiration processes etc which are mentioned in allopathy by various researches.

2. **Pharmaceuticals** - various types of Antibiotics, Anti Hypertensive drugs, Anti Diabetic drugs, Pain killers, Anti Histamine, Anti Toxic etc., pharm cogenesis is done by applying various research methodologies.
3. **Diagnostic** - various diagnostic methods like Blood lab investigation, Urine lab investigation, X-ray, USG , PET (Positron Emission Tomography), CT scan (Computerized Tomography), MRI (Magnetic Resonance Imaging), ECG, Eco-Cardiograph, FNAC, Angiography etc. diagnostic procedure and method are invented their scientist and researchers step by step.
4. **Therapeutic**
   (a) Various surgeries of various systems for e.g. open heart surgery, knee replacement surgery, plastic surgery, brain surgery etc.
   (b) Various treatment procedures for e.g. I.V., Ventilators etc. procedures are used in allopathy.

(B) Research in Homeopathy

1. Homeopathy develops on principles of similia - similibius counter.
2. By using mono remedy and poly remedy they cured various chronic diseases.
3. Over the years step by step with observation and experiments it went on developing step by step.

So Ayurveda, Allopathy, Homeopathy etc. health sciences are going in developing stage due to various researches related to those health sciences.

(2) Evidences of Researches in Ayurvedic classics

The Ayurveda is the ancient system of medicine whose antiquity goes back to Veda. Though ayurveda has divine origin, it shows glimpse of research. Word research itself implicates the meaning search again & again, which reflects that research.

Research is continuous process & outcome of it should maintain gold standard throughout. If it is not so, it should change timely i.e. must follow "Yuganurup Sandarbho."

There are some examples of researches mentioned in ayurveda:—

1. Sharir Kriya : Entire processes of digestion (sthoolpachana) has been

divided into three stages known as Avasthapaka where time required for digestion ingested food & lakshana (symptoms) developed during the processes of digestion has been mentioned.

Co-relation between Agni (digestive power) & digestion process (Annapachan) is described very well.

2. Sharir Rachana : Number of bones (Asthi) are mentioned.

Koshtasharir, Marmsharir is also described in Rachanavidnyana. Along with this Marmasthana, Marmaghatlakshana & their types are mentioned.

Different organ with their anatomical description function is given in Rachanasharir in detail.

3. Rasa Shastra : Numerous medical preparations (Bheshaja Kalpana) likeswarasa (for fresh juice), kwatha (decoction), awleha, malahar (ointment) etc have been described vividly. Today also medicine is prepare the following the same guidelines mentioned previously.

Use of various minerals, gems (Ratna-Upratna) for medicine preparation is mentioned.

4. Agadtantra : Sign & symptom of scorpion bite (vrischik Damsha), snake bite (Sarpa damsha)etc along with treatment is mentioned.

5. Kayachikitsa : Numerous diseases their sign & symptoms, treatment aspect in well elaborated in our treaties which can be correlated with diseases found today.

6. Dravyaguna : Aahardravya (food items) gana (group of medicines) are mentioned.

Various vishdravya are used as medicine judiciously.

7. Shalyatantra : Yantra & Shastra (instrument) mentioned by ayurveda are near about similar to today's instruments with minor changes.

Dahan karma is used to conduct previously. Today also in the form of cauterization it exists.

8. Streeroga : Description regarding Raja, Artava etc. get matched with today's modern Gynecology & obs.

Contra ceptive measures (medicines) are still followed as mentioned by our ancient aacharyas.

9. Shalakyatantra : Diseases of Eye, Ear are mentioned precisely in ayurveda which are observed till today.

10. Panchkarma : To remove impure blood (Dushta Rakta) from our body, leech application (Jalaukavcharan), Vein section (Siravedha) was conducted. Same principle is followed through bloodletting process.

11. Swasthavritta : Pranayama yoga is described very well & is practiced today also by healthy people & diseased one.

12. Balroga : Various balroga with their management is mentioned in our text. It is useful today also.

13. Samhita : Various disorders with their names, types, sign & symptoms along with treatment are mentioned which serves as guidelines today also.

Types, properties of madhu (honey) & their effect (mode of action) on body is also described minutely.

Panchdnyanendriya (5 sense organs), panchkarmendriya & Mana (mind) is elaborated very well in the text.

14. Rog Nidan : Processes of disease development through Shat Kriya Kala (six stages through which disease get established) is mentioned in the text.

Without any instrumentation the urine examination, stool examination, breast milk examination is described previously.

QUESTION :

1. Explain brief historical background of research in ayurveda
2. Write evidences of researches in ayurvedic classics.

• • •

Chapter 2

ANUSANDHAN

CHAPTER OVERVIEW
- Definition of Anusandhan
- Synonyms of Anusandhan
- Definition of Research
- Aims & objectives of research in ayurveda

अनुसन्धान

कार्यकारणभावस्य द्रव्याणां गुणकर्मणोः ।
परीक्ष्य स्थापनं सम्यगनुसन्धानमुच्चते ।।
अन्वेषणं पर्येषणं गवेषणमथापि च।
शोध इत्यादिपर्यायेः लोके समभिधीयते।।

–आचार्यप्रियव्रत शर्मा

अनु-अनन्तर

सन्धान-शोध / सम्यकज्ञान

व्याख्या : खोज की वह प्रक्रिया है जिसके द्वारा वस्तुओं और विचारों के बीच स्थित कार्य कारण भाव का अध्ययन कर उसे स्थापित किया जाता है।

Synonyms

(1) **अन्वेषण–अनु–अनन्तर / एषण–इच्छा।**

इष्यते उन्वष्यतेसाध्यतेनमेत्येषणा। च.सु.11/1(चक्रपाणि)

खोज की इच्छा निर्माण होना।

(2) **पर्येषण–(परि–चारो ओर / एषण–इच्छा।**

वस्तु का सर्वांगिण अध्ययन करना।

(3) **गवेषण–**

(i) **गोभिः इन्द्रियाः एषणं।**

अमरकोष : इन्द्रियो से वस्तुओं का परीक्षण करके उनकी खोज करना।

(ii) **गवामिव एषणमं।**

जिस तरह सरव: जंगलो में भटकी गायों की खोज करता है उसी प्रकार लुप्त वस्तुओं या विचारों की खोज करना गवेषण कहलाता है।

(4) **शोध**

पुराने विचारों का शोधन कर नवीन युक्ती युक्त विचार श्रृंखला स्थापित करने के कारण अनुसन्धान के लिए शोध शब्द का प्रयोग होता है।

अत: वस्तुओं का यथार्थ ज्ञान प्राप्त करने के लिए खोज की प्रक्रिया ही अनुसन्धान है।

DEFINITIONS OF RESEARCH

1. According to WHO : Research is a quest for knowledge through diligent search or investigation or experimentation aimed at discovery and inter-relation of new knowledge.

2. According to C.A.Moser : Research is a systematic investigation to gain knowledge about phenomenon and problems.

3. According to E. Bogardue : Research is an investigation of underline process, operative in lives of a person who are in association.

4. According to P.V.Young : Research is scientific undertaking which by means of logical and systemized methods, aims to discover the new facts or to verify old facts and to analyze their sequences, inter-relation, casual explanations and natural terms which given them.

Aims and objectives of Research in Ayurveda : Without clear aims and objectives, a research project can start on wrong foot, ending into fruitless and non-conclusive reports. The major problem faced by an Ayurvedic Research worker is the unavailability of standard protocol for basic Research.

Ayurvedic literature explains Aims as proper knowledge or conclusion.

(I) Aims of Ayurvedic Research

(1) Desire to get absolute knowledge

(2) Desire to better standard living

The main aim is to find out the hidden truth and the truth which is not been discovered yet.

(II) Objectives

1. धातुसाम्य क्रियाचोक्तातंत्रस्यास्य प्रयोजनं। च.सु. 1/53
2. स्वस्थस्य स्वास्थ्य रक्षणं आतुरस्य विकारप्रशमनंच्।
   Maintenance of Health and treatment of disease is the main objective.

3. To understand an untoward phenomenon and to find out the ways to avoid the same.
4. To understand a useful phenomenon and to find out methods to achieve continuation and improvement in it.
5. To investigate the principles of ayurveda and find out the ways of its practical application to suit it to modern lifestyle.
6. To prove Ayurvedic principles from modern point of view.
7. To elaborate and explain Ayurvedic principles for establishing Ayurveda as global science.
8. To study the cause – effect relationship for the proper understanding of Ayurvedic concept
9. To search for hidden truth under the observed phenomenon.
10. To find solution for current problem like autoimmune disease, sound and air pollution
11. To find out the truth or focus the light on old facts, concepts and practices. They should be first challenged and doubted. Critically re-examined in modern life. Then accept or reject it only after convincing evidence is found.
12. In this, some points needs elaboration because many topics have short descriptions or less classification.

अव्यक्तअनुक्तलेशोक्तसंदीग्धार्थप्रकाशक: ।

Here some points even raise doubts. All these points can be made clear and under by an intelligent persons with the help of techniques तंत्र युक्ती exposition or scientific presentation.

QUESTION :

1. Explain Synonyms of anusandhan
2. Write the definitions of research
3. Write the aims & objectives of research in ayurveda

• • •

Chapter 3

SCOPE, NEED AND IMPORTANCE OF RESEARCH IN AYURVEDA

CHAPTER OVERVIEW

- Scope of Research
- Need or Research
- Importance of Research

1. SCOPE OF RESEARCH IN AYURVEDA

Introduction

Research is a lifelong passion. It is driven by a restless curiosity that is refined by meticulous attention to the minutest details. Research necessitates high self-esteem, advanced skills, robust knowledge and humility. The last quality bespeaks of the fact that science is a collective global enterprise. All this requires years of apprenticeship and training. Currently we do not have adequate numbers of scientists with dual competences in both Ayurveda and basic sciences. Without such a critical mass, it is unrealistic to expect high quality research.

It is desirable to fill such a lacuna by creating a new genre of Vaidya-Scientists. This is not meant even remotely to imply that Vaidyas are not scientists or that physicians cannot do science. Vaidya-Scientist is proposed as a category like physician-scientists or clinical researcher, which are rare species even in modern biomedicine.

Physician-scientists need in-depth training in medicine as well as research methodology. Similarly, Vaidya-Scientists, who opt for academic or industrial Ayurvedic research, would need a special niche in institutions, and careful nurture and support by research councils.

The scope for research in Ayurveda has expanded significantly INTO 3 themes1) Theoretical research 2) Experimental research and 3) Clinical research. The subsets of these categories have also been described. Theoretical research in fundamental principles of Ayurveda has a potential for major paradigm shifts in bio-medicine. For example, understanding the scientific correlates of Gunas, Doshas, Dhatus and Malas would

better define the prognostic sets and subsets in the taxonomy of disease. The uniqueness of concepts like Pragnyaparadha, Ritucharya, Pathya-Prakriti and Kayakalpa would be an insistent invitation for research in life sciences. Experimental research in Ayurveda, while respecting the Baconian method, should transcend it by innovative research paths. The latter should not neglect the modes of evidence in Ayurveda viz. Apta, Pratyaksha, Anumana, Upamana and Yukti. Clinical research in Ayurveda offers the most fruitful domain for new knowledge. As Ayurveda is currently practiced by more than 700,000 vaidyas, the realm of Ayurvedic bed-side observation is rich and relatively unexplored. The emphasis in clinical research has to be on evidence based, diverse in modes of validity: consensual, congruent and concurrent. There is also an urgent need for translational research in Ayurveda.

Recently several interesting paths have been evolved for research and development in Ayurveda. For natural product drug discovery, Ayurvedic Pharmacoepidemiology and Reverse Pharmacology have been well initiated and partly followed up. Ayurvedic Pharmacoepidemiology can provide rich data on drug utilization, safety and novel activities. Reverse Pharmacology then make use of such data to embark on its three stages: experiential, exploratory and experimental. The Reverse Pharmacology path has led to several hits of drug-like activities in medicinal plants. Leads have targeted mechanisms with well-defined plant extracts. As a result, plant principles have been identified as drug candidates with good safety and efficacy. Successful examples of this path include: Kutaki (Picrorhizakurroa) for viral hepatitis, Atmagupta (Mucuna pruriens) for Parkinson's disease, Haridra (Curcuma longa) as a cancer-preventive, Amruta (Tinospora cordifolia) as an immune modulator, and Ashwagandha (Withania somnifera) in anxiety neurosis.

2. NEED OF RESEARCH IN AYURVEDA

Research should be a process that converts data into information, information into knowledge and knowledge into wisdom. This is like transforming milk into ghee. It should be more balanced comprehensive, and equally emphasizing in the literary field, experimental and clinical research. It should be able to impact the fields of academics, pharmacy and practice in a profound way. Present day Ayurvedic researches are failing in this aspect as they are unable to disseminate the knowledge gained from the exercises. Neither has the Ayurvedic teaching changed in the last 50 years nor have the textbooks been enriched with new researches.

Predominantly, drug researches done in the field of Ayurveda in the last six decades have not enriched the Ayurvedic understanding or Ayurvedic concepts; however, these researches have created a better understanding of Ayurveda by the modern medical fraternity. The researches done in the last 60 years on Herbal Pharmacology have led confirmation of few concepts like Reverse Pharmacology and use of whole crude drugs in place of isolation of fractions for clinical trials. These leads have changed the mindset of researchers on herbal medicine. In the last decade, lot of interest has been generated in the medical world regarding Ayurveda and other traditional medicines. However, all these efforts lead to the enrichment of the knowledge of the modern medicine and inclusion of some Ayurvedic herbs in modern MateriaMedica.

Since the last few years, it has been felt that there is a great need for a separate research methodology for Ayurveda and traditional medicines. WHO made an attempt in this regard in 2000 (WHO/TRM guidelines). However, this document has strongly recommended the exhaustive training of manpower of Ayurvedic institutions for research methodology. But the Government of India could not invest in the area, leading to production of non-qualified researchers in Ayurvedic Institutions who are supposed to be the backbone of Ayurvedic researches.

The clinical trials in Ayurveda are needed for–

a. Revalidation of facts enumerated in Ayurvedic classics leading to the explanation of fundamental principles;
b. To find out better treatment modalities for the existing diseases and for newer diseases;
c. To standardize the treatment procedures scientifically and
d. To establish dose, duration, indication and side-effect profile of any given drug.

Hence, if Ayurveda has to provide the much needed support to the modern medicine in the management of the diseases, it requires more research in the areas of fundamental principles and diagnostic tools in place of drugs. For the purpose the modern research, methodology is not suitable and there is a need for a paradigm shift in the research methodology for Ayurveda. Though most of the renowned workers advocate the same,

when it comes to practice it is not acceptable.

3. IMPORTANCE OF RESEARCH IN AYURVEDA

Research is the systematic investigation and study of materials and sources to establish facts and reach new conclusions, so it shapes people's understanding of the world around them. Through research findings, psychologists are able to explain individuals' behaviors, including how people think and act in certain ways. This helps to determine disorders and their impact on the person and society, thus developing appropriate treatments to improve the individual's quality of life.

Research has led to the introduction of new medical treatments and cures that have helped counter several diseases, thus increasing human life expectancy. It is now possible to live 10 years longer than in the 1960s and 20 years longer than in the 1930s. Causes of early deaths and crippling vitamin deficiencies have also been identified by progress made in the medical field through research.

What are the purposes of research? Patton (1990) pointed out the importance of identifying the purpose in a research process. He classified four types of research based on different purposes:

1. *Basic Research*: The purpose of this research is to understand and explain, i.e. the research is interested in formulating and testing theoretical construct and propositions that ideally generalize across time and space. This type of research takes the form of a theory that explains the phenomenon under investigation to give its contribution to knowledge. This research is more descriptive in nature exploring what, why and how questions.
2. *Applied Research*: The purpose of this research is to help people understand the nature of human problems so that human beings can more effectively control their environment. In other words, this type of research pursues potential solutions to human and societal problems. This research is more prescriptive in nature, focusing on how questions.
3. *Evaluation Research* (summative and formative): Evaluation research studies the processes and outcomes aimed at attempted solution. The purpose of formative research is to improve human intervention within specific conditions, such as activities, time, and groups of people; the purpose of summative evaluation is to judge the

effectiveness of a program, policy, or product.

4. *Action Research*: Action research aims at solving specific problems within a program, organization, or community. Patton (1990) described that design and data collection in action research tend to be more informal, and the people in the situation are directly involved in gathering information and studying themselves.

QUESTION

1. Write the scope of research in ayurveda
2. Explain need and importance of research in ayurveda

• • •

Chapter 4

TYPE OF RESEARCH

CHAPTER OVERVIEW

- **Classification of Research**
  1. Descriptive & Analytical Research
  2. Fundamental & Applied Research
  3. Qualitetive & Quantitetive Research
  4. Conceptual & Emoirecical Research
- **Other Type of Research**
  1. Motivational
  2. Attitude or Opinion
  3. Action
  4. Longitudinal
  5. Serendipity
  6. Medical
  7. Epidemiological

CLASSIFICATION OF RESEARCH

1. Descriptive	vs	Analytical
2. Fundamental	vs	Applied
3. Qualitative	vs	Quantitative
4. Conceptual	vs	Empirical

(1) Descriptive and Analytical Research-

(a) Descriptive Research – It includes surveys and fact finding enquiries of difference kind. The major purpose of this is description of state of affairs as it exists at present.

Characteristics

1. No control over variables.
2. Only report what happened and happening. *Ex* – The frequency of shopping, the preferences of peoples etc.

The methods used in this type are the survey methods of all types i.e. comparative, co-relational method etc.

(b) Analytical Research – Researcher has to use facts or information which is already available and analyze these to make critical evaluation of the material.

(2) Fundamental and Applied Research

(a) Pure or Fundamental Research – It is also called as basic or pure research.

It mainly concern with generalization and with formulation of theory. It is nothing but gathering knowledge for knowledge.

It aims at finding out new knowledge which has more / less theoretical orientation use.

With the help of this type of knowledge (Research), Researcher throws light on the past knowledge, which is already given but he can also give new direction or study the situation by difference manner.

Ex. – In ayurvedic literature especially in fundamental concepts of Sharir (i.e. Anatomical and physiological aspect of human body) various terms are not simplified i.e. Kloma, Kala etc.

So with help of fundamental Research we can study these aspects and draw a conclusion which may help to simplify the understanding process of these terms.

The literature Research is basic Research module which is almost common in other Research module.

(b) Applied Research – It is also called as Action Research.

Aim

1. To finding a solution for an immediate problem facing a society or any organization.
2. This Research is aimed at a certain conclusion.

In this materialistic age, more and more emphasis is being placed on practical utility of knowledge.

The knowledge without practical relevance always gets least priority and considers being secondary.

Although the basic Research gives new aspects of knowledge, the Applied Research utilizes this knowledge for practical purpose.

This Area of this research has potential to identify the problem.

Ex. – facing a social or business problem.

Thus the central aim of Applied Research is to discover a solution for some pressing practical problem.

From the point of view of application, there are two broad categories of research: - pure research and - applied research.

Pure research involves developing and testing theories and hypotheses that are intellectually challenging to the researcher but it may or may not

have practical application at the present time or in the future. The knowledge produced through pure research is sought in order to add to the existing body of research methods.

Applied research is done to solve specific, practical questions; for policy formulation, administration and understanding of a phenomenon. It can be exploratory, but is usually descriptive. It is almost always done on the basis of basic research. Applied research can be carried out by academic or industrial institutions. Often, an academic institution such as a university will have a specific applied research program funded by an industrial partner interested in that program

(3) Quantitative and Quantitative Research

(a) Quantitative Research – Quantitative Research is based on measurements of Quantity or around. It is applicable to phenomenon, which can be expressed in terms of Quantity.

Ex – Infant mortality rate, difference ratio.

It has more value than Qualitative.

(b) Qualitative Research – If finding is based on statement or description called as Qualitative Research. It concern with Qualitative phenomenon.

It measures in terms of % (degree).

(4) Conceptual and Empirical Research –

(a) Conceptual Research – It is a research which is related to some abstract ideas or theory.

It is usually used by philosophers and thinkers to develop new concepts or to re-interpret existing one.

(b) Empirical Research – It is based on personal experience of researcher and is considered to be the most scientific.

It relies on experience of observation alone, often without due regard for system and theory.

It is data based research, and based on scientific observations and it is original and precise.

It comes up with conclusions which are capable of being verified by an observation or experiment, so it can be called as 'Experimental Type of Research.'

In such research the researcher must first provided himself with a working Hypothesis or go as to a probable result.

OTHER RESEARCH

(a) Motivational Research – It aims at discovering the underlying motive and desires.

Techniques used in this research are –

(a) Word associated test.

(b) Sentence completion test.

(c) Story completion test.

(b) Attitude or Opinion Research – Aims to find out how peoples feel about any particular subject.

(c) Action Research – Itislinked with applied research in terms of immediate utility.

As the name suggest in action Research aims at solving the problems which has already become the part of action plan.

It may be linked with object and it utilized to improve the situation of project.

In planning Centre, action Research is very important and used routinely.

(d) Longitudinal Research – The research is based on either the purpose of Research or time required to accomplish Research or on the basis of some another similar facts.

From the point of view of time, we can think of Research either as on time Research or longitudinal Research.

Research can be a field setting Research or laboratory Research or stimulation Research, depending upon the environment is to be carried out.

It can be classified as

1. Conclusion oriented.
2. Decision oriented.

(a) Conclusion Oriented – Here researcher is free to pick up a problem. He re-designs the enquiry as he proceeds and is prepared to conceptualize as he wishes.

i.e. (1) Free to pick up a problem.
(2) Can redesign enquiry as he proceeds.
(3) Prepared to conceptualize as he wish.

(b) Decision Oriented – Always need a decision maker and the researcher in this case not free to embark upon Research according to his own inclination.

(e) Serendipity – It is the occurrence of events by 'chance'; this concept can be stated by statistical concept of probability.

It tells how often a particular event takes place with how much certainty.

Biological events may occur due to chances of some triggering factor.

Research methodologies and statistics are the tool that helps in studying this occurrence and events.

Serendipity Research : Is mean with which ideas are generated.

The Researcher get inspiration from such ideas and plans of study from such idea, he get transformed to discovering.

Ex -

1. Coronary catheterization – was discovered as method when radiologist at Cleveland clinic accidently injected Radio contrast into coronary Artery instead of ventricle.
2. Pap smear Test – discovered by Dr. Georgeons Papanikolous when he was investigating cycle of cell in vaginal fluid of CA patient.
3. Vaccines –
4. Antivirals –
5. Penicillin –
6. Vitamin A – action on wrinkle. Previously scientists gave vitamin A to patient of Acne but patient becomes wrinkle free.
7. Libido enhancing effect of L-dopa – L dopa when given for Parkinson Disease causes increase in libido.

(f) Medical Research – It can be classified as –

(a) Drug Research
(b) Clinical Research
(c) Epidemiological Research etc.

(g) Epidemiological Research – Epidemiology is the study of distribution and determination of health related states or events in specified population and application of study to control Health problem.

Types

(a) Observational
(b) Experimental.

(a) Observational – It is a descriptive and analytical study.

(b) Experimental – Experimental Research studies are randomized control trialsorclinical trials. This includes shifting field trials which has community interventions with Healthy peoples as a unit of study.

Although these are the difference types of Research studies, they can't be separated into watertight compartment.

QUESTION

1. Explain the classification of research
2. Write short note on pure & applied research
3. Write short note on observational & interventional research
4. Write short note on qualitative & quantitative research

• • •

Chapter 5

STEPS IN RESEARCH PROCESS

CHAPTER OVERVIEW

- Selection of Research Topic
- Review of Literature
- Hypothesis
- Aims & Objectives
- Material & Method
- Observation & Results

RESEARCH PROCESS

(1) Selection of Topic : First step is to identify or to select or to develop a research problem or a question.

Research problem is the one which require a researcher to find out the best solution for given problem.

Generally it refers to some difficulty that is either theoretical or practical solution of that problem and researcher wants to be obtaining the solution.

The two basic things are required for selection of problem.

1. That particular research must be adding something in knowledge.
2. Research work must be useful and priority to public health.

Avoid – Controversial subjects.

Too narrow or too vague problems.

Requirements : Subject must be familiar and feasible.

Researcher must understand – cost, time factor, qualification, training for his research.

Define Research problem and Statement : Topic of researchmust convey in specific terms, that what scientist is studying.

Scientist should be defined and clear on his undertaking.

The Statement of problem should not tell the whole purpose but must give rough idea.

Ex. – दशमुल त्रिवृत्तसिद्धतैल मात्रवस्ति in योनिव्यापद्

(2) Review of Literature : Researchstudy never conducted in vacuum so

the efforts should be made to find out what has already done and what is still to be done.

This gap should be marked and attempt should be made to bridge the gap.

If this step is not followed properly, duplication of the work is quite possible.

A Review of literature gives both theoretical as well as methodological directions.

For Ayurvedic research, student should review and it is better to have look on various periodicals and journals.

(3) Formulation of Hypothesis : It means mere assumption or some suppositions to be proved or disproved, but for Researcher it is a formal question that he intends to resolve.

It is a suggested solution to a problem.

Hypothesis is a proposed explanation based on limited evidence used as starting point for further investigation. In short it is foundation of future study.

State a Hypothesis : After making a problem and purpose clear and literature on previous work is reviewed. You have to precisely start with an assumption either positive or negative.

Ex. – Iron by intra muscular route is not more effective than by orally.

So test this Hypothesis by doing an experiment and accept or reject the Hypothesis.

Characteristics :–

1. It must be clear and precise.
2. It should be capable of being tested.
3. It should be limited in scope, must be specific.
4. It should start with simple fact or terms without changing its significance and should complete with time.
5. It must explain the facts that give rise to need for explanation.

Good Hypothesis Qualities

(1) Must be adequate answer to specific problem.

(2) Should be logical and simple.

(3) Should be expressed in quantified from.

(4) Must be variable.

(5) Must be started in way to allows it to be refused.

(4) Aims and Objectives : Formulating aims and objectives for your research studies helps sculpt and guide your work after you've decided on a topic. While your aims give your research thematic and theoretic direction, objectives give concrete steps on how to manifest those concepts and theories. A strong topic begins with an area you're interested in researching, then focuses on a need, problem or unexplored issue in that area, usually from a particular perspective or approach.

Aims : A research proposal's aims are statements that broadly point out what you hope to accomplish and your desired outcomes from the research. Aims focus on long-term intended outcomes -- your aspirations in reference to the research. They're typically not numbered in a research proposal. Strong research proposals set out only two to three aims. Each aim may have several objectives associated with it.

Objectives : Objectives lay out how you plan to accomplish your aims. While aims are broad in nature, objectives are focused and practical. They tend to pinpoint your research's more immediate effects. They include a list of practical steps and tasks you're going to take to meet your aims. Objectives are typically numbered, so each one stands alone. Each objective must have a concrete method set out. If you're having trouble developing this, writing out a research time-line before defining your objectives may help.

What They Should Do : Both aims and objectives should be brief and concise. They must be interrelated. Each aim should have one or more objectives describing how that aim should be met. Aims and objectives should both be realistic goals and methods with respect to what resources you have available and the scope of research. Don't choose something too broad, for instance, or that would take much more research time than you have. Aims and objectives also let your advisers know how you intend to approach a given subject and how you plan to get access to subjects, goods and services, samplings and other resources. They also provide plans for dealing with ethical or practical problems you may encounter.

Things to Avoid : When you're writing your aims and objectives, don't make them too broad or vague. Aims are more general than objectives,

they shouldn't be over-optimistic or unrealistic about what you want to achieve. They shouldn't repeat each other or be simple lists of items related to your research. Stay away from focusing content on your research site or job. Stick to the specific aims and objectives of your actual research study. Check that your aims and objectives contain methods that can sustain what you hope to find. If they don't match, rethink the methods or aims and objectives of your research.

(5) MATERIAL & METHODS

Introduction : The methods and methodology section of the research proposal provides an overall plan of how the research objectives will be achieved. It includes a description of the methods and the methodology. Methods are the techniques or procedures used to gather and analyze data or to make a piece of art. Methodology refers to the 'the strategy, plan of action, process or design lying behind the choice and use of particular methods, and linking the choice and use of methods to the desired outcomes' (Crotty, 1998:3). Methodology includes not only the practical aspects of the research such as method and action plan, but also the philosophical and theoretical perspectives of the research.

Methods : Before going on to consider how to write these sections in more detail, it might be useful to clarify what is meant by the terms methods and methodological rationale, and why theory or philosophy becomes important in social and human research:.

Methods are the physical activities that researchers undertake to gather data. Some common research methods include:

- Document analysis
- Textual analysis
- Policy analysis
- Interviewing
- Survey
- Experiment
- Observation
- Statistical analysis of public records
- Art making practice
- Documentary or film making.

(6) OBSERVATION & RESULTS

Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals.

Survey weights often need to be applied to the data to adjust for the sample design.

Results from probability theory and statistical theory are employed to guide practice.

In business, sampling is widely used for gathering information about a population.

Once the instrumentation plan is completed, the actual study begins with the collection of data. The collection of data is a critical step in providing the information needed to answer the research question. Every study includes the collection of some type of data—whether it is from the literature or from subjects—to answer the research question. Data can be collected in the form of words on a survey, with a questionnaire, through observations, or from the literature. In the obesity study, the programmers will be collecting data on the defined variables: weight, percentage of body fat, cholesterol levels, and the number of days the person walked a total of 10,000 steps during the class.

The researcher collects these data at the first session and at the last session of the program. These two sets of data are necessary to determine the effect of the walking program on weight, body fat, and cholesterol level. Once the data are collected on the variables, the researcher is ready to move to the final step of the process, which is the data analysis.

Analyze the Data : All the time, effort, and resources dedicated to steps 1 through 7 of the research process culminate in this final step. The researcher finally has data to analyze so that the research question can be answered. In the instrumentation plan, the researcher specified how the data will be analyzed. The researcher now analyzes the data according to the plan. The results of this analysis are then reviewed and summarized in a manner directly related to the research questions. In the obesity study, the researcher compares the measurements of weight, percentage of body fat, and cholesterol that were taken at the first meeting of the subjects to the measurements of the same variables at the final program session.

These two sets of data will be analyzed to determine if there was a difference between the first measurement and the second measurement for each individual in the program. Then, the data will be analyzed to determine if the differences are statistically significant. If the differences are statistically significant, the study validates the theory that was the focus of the study. The results of the study also provide valuable information about one strategy to combat childhood obesity in the community.

As you have probably concluded, conducting studies using the eight steps of the scientific research process requires you to dedicate time and effort to the planning process. You cannot conduct a study using the scientific research process when time is limited or the study is done at the last minute. Researchers who do this conduct studies that result in either false conclusions or conclusions that are not of any value to the organization

QUESTION

1. Write the research processes & its importance
2. Write importance of-
   (a) Selection of topic of research
   (b) Review of literature
   (c) Hypothesis
   (d) Aims & Objectives
   (e) Material & method
   (f) Observation & results

Chapter 6

ROLE OF PRAMANA AS A RESEARCH TOOLS

CHAPTER OVERVIEW

- Aptopadeshapramana
- Pratyakshapramana
- Anumanapramana
- Yuktipramana

INTRODUCTION

Ayurveda is a comprehensive system of health care embarked thousands of years ago and is successfully being practiced even today. Ayurveda is the science of healthy and diseases free life. In our classics the very basic definition of health is defined as the equilibrium state of Dosha, Dhatu, Agni and Mala as well as the pleasant state of Aatma, Indriya and Manas. The science of Ayurveda is based on sound principles of diagnosis which requires no clinical or costly investigations which are beyond the reach of the common and less privileged persons.

Ayurvedic science mainly aims at removing the root cause of the disease rather than simply trying to suppress the symptoms. To find out the root cause of a disease and to plan the appropriate treatment Acharyas espoused the fourfold tools of knowledge referred to as Pramana. Real knowledge or the actual happening is known as the Prama or the truth, whereas the means by which the truth is checked is known as Pramana.

In other words it can be said that Pramana is the means that detects the difference between truth and lie, actual and virtual. Charakacharya considers four Pramanas as the means of Pariksha karama and those are pivotal in diagnosis, prognosis and treatment, namely Pratyaksha, Anumana, Aaptopadesha and Yukti.

Research can be considered as a journey from an idea i.e. stating or formulating a research problem to a valid conclusion. In ayurveda, the pramana vidnyaya can be considered as a scientific tool of research. These pramanas are not only useful in Roga and rogi pariksha but also it is useful to understand the principles of ayurveda thoroughly. Charakacharya has advised to adopt these pramanas as a tool to get clear and entire knowledge of ayurveda.

Pramana can be considered as one of the ancient method of research in developing research methodology in ayurveda. According to

Charakacharya the things of universe such as sat (Existent) or asat (non-existent) can be investigated by these four Pramanas and the same are helpful in swasthyarakshan, roga pariksha, rogi pariksha, dravya pariksha and chikitsa.

They are –

1. Aaptopadesh Pramana
2. Pratyaksha Pramana
3. Anumana Pramana
4. Yukti Pramana

These four pramanas (Aaptopadesh, Pratyaksha, anumaan and yukti) are very much useful at all steps of research such as planning of research work, executing the plan of research and finding a conclusion.

As we know the word Research refers to a search for knowledge. It is nothing but a planned program which is used to solve problem & creates a generally applicable new knowledge.

In ayurveda the word praman is derived from prama which means valid or correct knowledge i.e. knowledge of science which has already been explored by various scientific methods.

The critical scientific approach of ayurveda is evident from its Pramana vidnyan. Thus pramanas are the scientific methods to acquire valid knowledge.

Among all pramanas, aaptopadesh gives existing knowledge of science, while pratyaksha and anumana gives knowledge after examination. Aaptopadesh is considered as the primary one, as it forms the first source of information about ayurveda or any other science. It forms a firm basis for all theoretical knowledge in all aspects.

1. AAPTOPADESH PRAMANA

Term aaptopadesh is generally confined only with the individual or personality. Aapta is not merely an individual but also it includes written documents like vedas, samhitas, different manuscripts etc. Now a day various research journals, different scientific websites can be taken or consider as Aapta.

Before any research or proposed study, we follow the step i.e. review of literature which can be correlated with aaptopadesh. It helps to avoid repetition of work or research. Also it helps to get fundamental understanding about any research problem and on the basis of previous research work done, one can plan further or next study as further scope of research.

As we know that, Aapta vachana is universal truth. The Siddhant which is put forth or put forward by examining the fact thoroughly, which forms fundamental store for any research work.

In short aapta means yatharth vakta i.e. precise speaker and the things which are said by these aaptas are considered as Aaptopadesh. The upadesha given by aapta is considered as pramana. So let us know the qualities of aapta or to whom we can call as aapta:-

- Those who were free from raja and tama (manas doshas)
- They have knowledge of all three kalas i.e. past, future and present.
- They have un-doubtful knowledge etc.

As aacharya said that the aapta is merely not only an individual but it also can be vedas, samhitas etc. so on that context we can say that the researches which are proven also considered as aapta for new researcher.

A researcher who want to research on one topic, get idea about that topic with the help of previous researches. Hence the work which was previously proved scientifically will become aapta for that new research or study.

Researcher should acquire the existing knowledge & training in physical & mental skills which is necessary to do the activities implied in research. The existing knowledge of science can be acquired through Aaptopadesh.

As our ancient aacharyas said that the information which was written or told in our samhitas is approved by the ancient scriptures. And since their benefits are perceptible, these should be administered as it is without any discussion.

Some says that the ayurvedic medicines don't have any evidence base. But as we know that ayurveda science is available since ancient times, so it has been based on evidences.

The various formulations which are given in our text are still useful and effective. These formulations with their matra or pramana are the evidences of research of that time.

In some cases, things which are mentioned in our samhitas have not clear idea about them for example the concept of kloma. In ayurvedic texts the concept of kloma is still not get cleared. So to understand some of these facts or concepts we have to do research to reveal them with the help of this aaptopadesh.

Aaptopadesh is helpful in Roga pariksha and Chikitsa. Before arriving at a conclusion with regard to the diagnosis or treatment of a disease, it is essential to contemplate on the teachings of the seers. For an Ayurvedic physician today, it would mean considering the information available in the major treatises of Ayurveda such as Charaka Samhita and Sushruta Samhita to augment diagnosis, and to implement proper treatment measures.

Thus, Aaptopadesh is considered as the reliable source of existing knowledge, which suggest new ideas for research & also it helps in the development of research methodology.

IMPORTANCE OF APTOPADESHA PRAMANA

- In roga pariksha
- It becomes a base for Pratyaksha and Anumana Pramana.
- Nidana, roopa and chikitsa of any disease is known by Aaptopadesha.
- Without the knowledge of Aaptopadesha – pariksha or chikitsa is impossible.

2. PRATYAKSHA PRAMANA

The knowledge which is perceived by the sense organs & mind is called as Pratyaksha & the medium of it is called as Pratyaksha Pramana. Here knowledge is the result, while contact of the sense organs & the object is the medium of knowledge.

As we know, Charakacharya in vimanasthana said that the proper knowledge of a disease can be gained through three Pramanas-

1. Aaptopadesha Pramana
2. Pratyaksha Pramana
3. Anumana Pramana

Out of these, by using Pratyaksha Pramana- except Rasa, other senses i.e. shabda, sparsha, rupa & gandha can be perceived directly. Thus pratyaksha Pramana helps in Rogi and Roga Pariksha.

To make any research, one should collect the data first, then after that analysis and final conclusion will be drawn. So for the collection of data for any research, there are two main approaches:

(i) The Information which is required is already exist or available and only we have to extract it.

(ii) New and fresh information we have to collect.

Also as we know that, the types of data according to source are primary and secondary data. Among these, the data which is collected from primary sources are known as primary data. This can be achieved by observation or through questionnaire or by interview method.

These methods of collection of data more or less depend upon Pratyaksha pramana. Out of other methods of observation like experiments, survey and records, the experiment and up to some extent of survey method depends on this Pratyaksha pramana.

The Pratyaksha pramana is depends on five sense organs and mind, and this type of knowledge is gained directly. But it has some limitations which were mentioned in our samhitas as Pratyaksha Badhak bhavas. Due to these factors or Pratyaksha Badhak bhavas, the pratyaksha knowledge is not achieved or gained. These limitations can be overcome by modern scientific equipment.

For example: - Ati-sukshma (very minute / small) particles or cells are not

seen by naked eyes, so here an obstacle comes for pratyaksha knowledge. So to overcome this problem microscope is used.

Also for aavarana i.e. due to covering one can't get direct knowledge of that structure, so to overcome this problem we can use x-ray, MRI or USG scan etc.

IMPORTANCE OF PRATYAKSHA PRAMANA

- In this way Pratyaksha helps in rogi pariksha and roga pariksha.
- For Darshana, sparshana and prashna pariksha - It is the base for the diagnosis and prognosis of a disease.
- To check the improvement in a patient pratyaksha is compulsory.
- Ashtavidh pariksha, Dashavidha pariksha, sthanik parikshan need mostly the support of Pratyaksha.
- It gives reliable information about physical & mental changes in patient.

3. ANUMANA PRAMANA

Anumana is indirect knowledge which is based on reasoning (tarka) or inference and the inference is based on prior perception (poorva dnyana). On the basis of direct perception inference of three kalas i.e. past, present and future will be drawn. Also there are three types of anumana –

(1) Sheshavata

It means inference of cause (karana) from effect (Karya) which is related to past.

For example – The undigested food vomiting can imagine adhyashana as its cause.

In terms of research this can be correlated with Case control study design and said to be retrospective study.

(2) Poorvavata

It means inference of effect (karya) from cause (karana) and it is related to future.

For example – Hrullasa in Chardhi and Jrumbha in Jwara disease.

In terms of research this can be correlated with cohort study design and said to be prospective study.

(3) Samanyoatodrushta

This type of anumana is related to present. For example - Inference of Agni by jaranashakti, Bala by vyayamashakti etc.

In terms of research this can be correlated with Cross sectional study.

Anumana helps in establishing the correlations between two factors or we can say it helps to establish cause – effect relationship (Karya – karana bhava sambandha) between two variables.

Anumana pramana is very essential in roga and rogi pariksha. Also it

helps to establish the cause – effect relationship according to ayurveda. As we know, in direct perception there are many obstacles i.e. pratyaksha badhak bhavas also the indirect knowledge can't be gained by pratykasha pramana. Hence to overcome this, the anumana pramana helps. It helps in getting knowledge and plays an important role in interpretation of results.

This interpretation of result is done by drawing the interference from collected data after doing its analysis or experiments.

Anumana Pramana is of two types –

1. Swarthanumana i.e. (it is for one's own self) &
2. Pararthanumana i.e. (it is used to provide knowledge to others).

(1) Swarthanumana

In this type of anumana, the efforts were made to get knowledge for that person himself.

(2) Pararthanumana

After getting knowledge by own, it should be demonstrated to others and for that purpose the 5 components of reasoning are stated. So with the help of these panchavayavi vakyas (5 components of reasoning), the inference is drawn is known as Pararthanumana.

These 5 components of reasoning are

1. Pratidnya
2. Hetu
3. Drushtant or Udaharana
4. Upanaya
5. Nigamana

So with the help of Pararthanumana, one can give knowledge to others which was got by himself previously and this can be correlated with writing the reports or results.

IMPORTANCE OF ANUMANA PRAMANA

- Anumana pramana is highly useful in areas of unavailability of sufficient information. In such a case, the physician is required to observe the available links and arrive at a conclusion.
- Anumana spans over three periods of time: the past, the present, and the future.
- Charaka elaborates further with discreet examples. This refers to an inference of the present situation based on an available clue.
- Anumana plays a major role in decoding the etiology of a disease and in diagnosis.

(4) YUKTI PRAMANA

Yukti comes from the root Yuj, which means to bring together, join. Yukti is the synonym of Yojna i.e intelectually planning.

Yukti is the Budhivishesha needed for perception of the common knowledge of the multi reasons. Yukti may be translated as analysis. The inferential knowledge that one arrives at after a careful introspection of various factors is defined as yukti.

The knowledge which sees the things produced by the combination of multiple causative factors is known as Yukti. It is firstly accepted and mentioned by Charak as a pramana. With the help of yukti pramana, we get the knowledge of three kaalas (three periods of time). It also helps to get knowledge of tri varga i.e. dharma, artha and kama.

Through proper presence of mind, intellect and co-ordination of the Atma, sense organs and their objects (artha/vishaya), one is able to perform all activities properly.

Yukti proves to be an important tool for diagnosis and treatment when there is involvement of multiple doshas, presence of confusing symptoms, and availability of few treatment options due to contradictions in the disease and constitution of the patient.

Yukti helps the physician to take decision based on his medical knowledge, practical experience, intelligence, observation and analytical skill to formulate an effective treatment. Kriyakala, agni, aushadhi sevana kala, matra etc. are known by Yukti pramana.

Yukti is considered to be an Anumana Pramana. But it is notable that Anuman is made by the means of Yukti and this is when it is considered as Yukti.

In medical science, if doctor knows which drug can be effective for a particular disease then the treatment becomes easy and effective. But sometimes we see that it's not fit for every situation i.e. one specific drug of that particular disease will not suitable for all patients, it's not useful to some patients due to variations in prakruti, bala, kaala, matra, age etc. So sometimes it's become very difficult to treat that disease. Hence to overcome those problems Charaka mentions this Yukti pramana.

With the help of this yukti pramana, if doctor plan for treatment by considering all these factors like age, kriya kaal, agni, bala etc. he will surely give better treatment which in turns cure that disease.

In same way, as far research concern, researcher must adopt this yukti pramana for his work or study. It helps him to come to final conclusion by considering all observational data, statistical analysis etc.

For the right application of Pratyaksha, Anumana and Aptopadesha. Yukti is the base. One may have extensive knowledge about a subject but without the Yukti it is of not much use.

Pratyaksha and Anumana are useful in diagnosis of the disease. Aptopadesha is mostly helpful in treatment aspect whereas the Yukti has importance in diagnosis, prognosis and treatment aspect. Yukti has clinical as well as philosophical importance. Yukti helps in maximum utilization of the available knowledge.

Most of the physicians gave importance to Pratyaksha, Anumana and Aptopadesha, but forget the clinical application of Yukti. When the first three Pramana's are applied by using Yukti it leads to the success.

Yukti is an important contribution of Charaka in the field of applied medicine. A closer look at all the Pramanas will bring to light the fact that they all fit in very well in today's scientific era. They stand a testimony to prove that true knowledge never changes with the passing time.

In fact all experiments being done today are based on Pratyaksha, Anumaana, Aptopadesha and mainly on Yukti Pramana through which the goals of human life namely Dharma, Artha, Kama and Moksha are achieved. Thus in present scenario there is a need to put more light on the concept of these pramanas.

IMPORTANCE OF YUKTI PRAMANA

- Yukti proves to be an important tool for diagnosis and treatment, when there is involvement of multiple doshas, presence of confusing symptoms, and availability of few treatment options due to contradictions in the disease and constitution of the patient. The decision taken by the physician based on his medical knowledge, practical experience, intelligence, observation and analytical skill to formulate an effective treatment.
- The yukti pramana plays major role in nidana & chikitsa. Ex – If the physician notices a particular group of signs and symptoms together he decides the course of disease & measures to be taken.
- The Kriyakala, agni, aushadhi sevana kala, matra etc. are known by yukti pramana.
- To confirm, evaluate and to treat a disease, physician need to perform clinical examinations of patient where aaptopadesha, pratyaksha and anumana are all very important components.

QUESTION

1. Write in details the role of pramanas as research tools
2. Importance of Anuman Pramana in Research
3. Importance of Pratyksh Pramana in Research
4. Importance of Aptodesh Pramana in Research
5. Importance of Yukti Pramana in Research

Chapter 7

ETHICS IN RESEARCH

CHAPTER OVERVIEW

- Selection of Research Top Ethical Aspects Related to Human
- Ethical Aspects Related to Animal Experiment
- General Information of Institutional Ethics Committee
- General Information of Animal Ethical Committee (AEC)

ETHICS

We are going through a time of profound change in our understanding of the ethics of applied social research. From the time immediately after World War II until the early 1990s, there was a gradually developing consensus about the key ethical principles that should underlie the research endeavor. Two marker events stand out (among many others) as symbolic of this consensus. The Nuremberg War Crimes Trial following World War II brought to public view the ways German scientists had used captive human subjects as subjects in oftentimes gruesome experiments. In the 1950s and 1960s, the Tuskegee Syphilis Study involved the withholding of known effective treatment for syphilis from African-American participants who were infected. Events like these forced the reexamination of ethical standards and the gradual development of a consensus that potential human subjects needed to be protected from being used as 'guinea pigs' in scientific research.

By the 1990s, the dynamics of the situation changed. Cancer patients and persons with AIDS fought publicly with the medical research establishment about the long time needed to get approval for and complete research into potential cures for fatal diseases. In many cases, it is the ethical assumptions of the previous thirty years that drive this 'go-slow' mentality. After all, we would rather risk denying treatment for a while until we achieve enough confidence in a treatment, rather than run the risk of harming innocent people (as in the Nuremberg and Tuskegee

events). But now, those who were threatened with fatal illness were saying to the research establishment that they wanted to be test subjects, even under experimental conditions of considerable risk. You had several very vocal and articulate patient groups who wanted to be experimented on coming up against an ethical review system that was designed to protect them from being experimented on.

Although the last few years in the ethics of research have been tumultuous ones, it is beginning to appear that a new consensus is evolving that involves the stakeholder groups most affected by a problem participating more actively in the formulation of guidelines for research. While it's not entirely clear, at present, what the new consensus will be, it is almost certain that it will not fall at either extreme: protecting against human experimentation at all costs vs. allowing anyone who is willing to be experimented on.

Ethical Issues : There are a number of key phrases that describe the system of ethical protections that the contemporary social and medical research establishments have created to try to protect better the rights of their research participants. The principle of voluntary participation requires that people not be coerced into participating in research. This is especially relevant where researchers had previously relied on 'captive audiences' for their subjects -- prisons, universities, and places like that. Closely related to the notion of voluntary participation is the requirement of informed consent. Essentially, this means that prospective research participants must be fully informed about the procedures and risks involved in research and must give their consent to participate. Ethical standards also require that researchers not put participants in a situation where they might be at risk of harm as a result of their participation. Harm can be defined as both physical and psychological. There are two standards that are applied in order to help protect the privacy of research participants. Almost all research guarantees the participant's confidentiality -- they are assured that identifying information will not be made available to anyone who is not directly involved in the study. The stricter standard is the principle of anonymity which essentially means that the participant will remain anonymous throughout the study -- even to the researchers themselves. Clearly, the anonymity standard is a stronger guarantee of privacy, but it is sometimes difficult to accomplish, especially in situations where participants have to be measured at multiple time points (e.g., a

pre-post study). Increasingly, researchers have had to deal with the ethical issue of a person's right to service. Good research practice often requires the use of a no-treatment control group -- a group of participants who do not get the treatment or program that is being studied. But when that treatment or program may have beneficial effects, persons assigned to the no-treatment control may feel their rights to equal access to services are being curtailed.

Even when clear ethical standards and principles exist, there will be times when the need to do accurate research runs up against the rights of potential participants. No set of standards can possibly anticipate every ethical circumstance. Furthermore, there needs to be a procedure that assures that researchers will consider all relevant ethical issues in formulating research plans. To address such needs most institutions and organizations have formulated an Institutional Review Board (IRB), a panel of persons who reviews grant proposals with respect to ethical implications and decides whether additional actions need to be taken to assure the safety and rights of participants. By reviewing proposals for research, IRBs also help to protect both the organization and the researcher against potential legal implications of neglecting to address important ethical issues of participants.

Ethical Aspects Related to Human : World Medicine Association (WMA) has developed the declaration of "Helsinki" as statement of ethical principles to provide guidance to physicians and other participants in medical research involving Human Subjects.

1. It is duty of physician to promote and safe guard the Health of people.
2. Declaration of WMA of Geneva binds the physician with words that are "The Health of my patient will be my first consideration".

   And physician shall act only in patient's interest when providing medical care which might have effect weakening the physical and mental condition of patient.
3. Primary purpose is medical research related to human is to improve prophylactic, diagnostic and therapeutic procedure and understanding of etiology and pathogenesis of disease.
4. It promotes the respect for all human being and protects their health and rights.

5. An Investigator should aware of ethical, legal and regulatory requirements for research on human subjects.
6. It is a duty of physician in medical research to protect life, health, privacy and dignity of human.
7. Appropriate caution must be exercised in conduct of Research which may affect the environment and welfare of animals used for Research must be respected.
8. Research protocol should always contain statement of ethical consideration involved.
9. Research should be conducted by scientifically qualified person and under supervision of clinically competent medical person.
10. Responsibility of human subject for Research must always rest medically qualified person and never rest on subject of research.
11. Physician should do all the investigations and medical care as per requirement but if any emergency occurs during research, it should be treated.
12. Subject must be volunteers and informed participants in the Research project.
13. Right of Research subjects to safe guard their integrity must always be respected.
14. Researcher must be take precaution and respect the privacy of subject and patient's information should not expose.
15. Physician must obtain an informed consent for Research project.

Ethical Aspects Related to Animal Experiment : The Medicinal and Scientific Research can't exist in isolation. Each new medicine has to be tested either on Animals or on tissue culture in laboratory before tested on human beings and eventually accepted as a treatment for any disease.

The Government constituted a Committee for the Purpose of Control and Supervision on Experiments on Animals in 1963 (CPCSEA)

Also an Animal Ethics Committee (AEC) has performed a major role in Ethics principles regarding Animals.

Ethical aspects or Principles are –

1. The proper health and hygiene of animals is requirednot for the sake of Animals but for the sake of Science.

2. Research project involving animal subjects should take place only when the aim of project is worthwhile and design of the project is such that, there is strong possibility that it will achieve the aim.
3. Breeding and experiments of animals can perform under the guidance of AEC.
4. The animals selected for experiment should be of an appropriate species and quality.
5. Investigators and other persons should treat animals with kindness and should take proper care of them.
6. Investigator should assume that all procedures which would cause pain in human beings may cause pain in other vertebral species.
7. Procedure that may cause more pain or distress should be performed under appropriate sedation, analgesic or anesthesia by taking help of veterinary doctors.
8. Normally the care of Animals should be taken under the supervision of veterinary doctor or a person having adequate experience of animal care.
9. Research strategy and small pilot studies are reviewed before committing an animal.
10. The unnecessary repetition of Animal procedures can be avoided.
11. In many areas of bio-medical science, in vitro methods are used.
12. Also Sometimes refinement alternatives encompass methods can be used and that method minimize the potential pain or distress and which enhances Animal well-being.

General Information of Institutional Ethics Committee : Ethical committee should be constituted and operated so that its task can be executed free from bias.

Ethical committee should have documented the policies and the procedures as a basis of its work.

The Documents should also state - how frequently the committee will meet and how it will interact with the investigator.

The Role of Ethical committee is to ensure the protection of rights and welfare of human subjects, participating in clinical trial.

Ethical committee looks after research process and keeps the safeguard rights intact of all possible living organisms.

Responsibilities

Safeguard rights, safe and well-being of trial subjects.

Review suitability of investigator and site facilities.

Review methods for obtaining and documenting consent.

Thoroughly review the protocol.

Review amount and method of payment to that subject.

Composition

At least 5-7 members are required.

At least 1 is non-scientific or non-medical.

At least 1 is from outside the institute.

1 experienced lawyer or judge.

Qualified and experienced scientist and clinician are taken.

Functions and Operation : The Functions are according to written operating procedure.

Committee should meet at least once in month to clear any pending proposals for ethics clearance, so that an investigator need not have to wait for longtime before starting their project.

Minimum quorum at meeting is must.

Experts can be invited as per requirement but they cannot vote.

Maintain the record at least for 3 years after completion of trail.

General Information of Animal Ethical Committee (AEC) : Role of AEC is to ensure that scientific use of Animals for scientific purposes. The welfare of these animals and incorporates the principles of replacement, reduction and refinement.

The chairperson of Animal Ethical Committee is Deputy Vise Chancellor (Research)

Functioning

1. Approve the Guidelines for care of animals that are breed, held and used for scientific purposes on behalf of university.
2. Monitor, acquisition, transportation, production, housing care use and fate of such animals.
3. Approve only those projects for which Animals are essential and justified.

4. To withdraw an approval for any project.
5. Authorize the emergency treatment.
6. Maintain a record of proposal and projects.
7. Committee usually meets 10 times in 1 year.
8. AEC shall inspect annually all campus and other facilities for keeping Animals used in approved protocols.

QUESTION

1. Importance of ethics in research
2. Write short note on Ethical aspect related to Human
3. Write short note on Ethical aspect related to Animal Experiment
4. Write short note on Institutional Ethical Committee

• • •

Chapter 8

EVIDENCE BASED MEDICINE AND SCIENTIFIC WRITING

CHAPTER OVERVIEW

- Evidence-Based Medicine (EBM)
- Research Article Structuring (IMRAD)

(1) Evidence Based Medicine : Evidence-based medicine (EBM) is a form of medicine that aims to optimize decision-making by emphasizing the use of evidence from well designed and conducted research. Although all medicine based on science has some degree of empirical support, EBM goes further, classifying evidence by its epistemologicstrength and requiring that only the strongest types (coming from meta-analyses, systematic reviews, and randomized controlled trials) can yield strong recommendations; weaker types (such as from case-control studies) can yield only weak recommendations. The term was originally used to describe an approach to teaching the practice of medicine and improving decisions by individual physicians. Use of the term rapidly expanded to include a previously described approach that emphasized the use of evidence in the design of guidelines and policies that apply to populations ("evidence-based practice policies"). It has subsequently spread to describe an approach to decision making that is used at virtually every level of health care as well as other fields, yielding the broader termevidence-based practice.

Whether applied to medical education, decisions about individuals, guidelines and policies applied to populations, or administration of health services in general, evidence-based medicine advocates that to the greatest extent possible, decisions and policies should be based on evidence, not just the beliefs of practitioners, experts, or administrators. It thus tries to assure that a clinician's opinion, which may be limited by knowledge gaps or biases, is supplemented with all available knowledge from the scientific literature so that best practice can be determined and

applied. It promotes the use of formal, explicit methods to analyze evidence and make it available to decision makers. It promotes programs to teach the methods to medical students, practitioners, and policy makers. The term "evidence-based medicine" was first coined and developed by doctors at McMaster University Medical School in the 1980s.

Evidence-based medicine is probably best understood as a decision-making framework that facilitates complex decisions across different and sometimes conflicting groups. It involves considering research and other forms of evidence on a routine basis when making healthcare decisions. Such decisions include clinical decisions about choice of treatment, test, or risk management for individual patients, as well as policy decisions for groups and populations

Evidence-based medicine has been particularly useful, however, for addressing questions that do not have intuitive answers, or those for which our impressions can actually cause more harm than good. The litany of medical procedures that everyone, at the time, felt to be beneficial, but which later proved to be either ineffective or harmful, is long. Thalidomide for pregnant women and internal mammary artery ligation for reducing ischemic heart disease are notorious examples, but plenty of others abound. By providing the data that are difficult to intuit, evidence-based medicine has helped to stem theoretically logical but potentially harmful decisions. For example, evidence-based medicine helped clinicians to realize the value of steroids for reducing respiratory distress inpremature babies, despite our long-held and understandable intuition, from centuries of puerperal fever, that steroids might be harmful in this context.

Other real examples of evidence-based medicine include the wider use of normal saline versus colloids for many kinds of resuscitation (again counterintuitive); methods for continuous professional development and education; and best-practice guidelines on asthma, heart failure, and cancer screening.

Perhaps more significant than evidence-based medicine's contribution to particular clinical questions, however, is how it has helped professional bodies to make decisions. By providing a fair, scientifically rigorous method for making best-practice decisions, evidence-based medicine has helped professional bodies to develop more transparent working practices

to establish guidelines and standards. In so doing, evidence-based medicine has proven timely, given the growing demand for professional accountability coupled with the technical complexity of healthcare decisions.

At a technical level, evidence-based medicine works by converting complex information from literally thousands of individual studies into user-friendly risk estimates. It involves a defined method with 4 main steps:

1. Defining a structured question about the target population, outcomes, and (usually) intervention or exposure;
2. Searching the published (and occasionally unpublished) literature for sources of data that might answer the question;
3. Appraising or evaluating the data for methodologic rigor and relevance to the question (typically, this process weeds out about 90% to 99% of studies; and
4. Describing and analyzing the resulting data to answer the question posed.

(2) Scientific writing of Research Article Structuring:-IMRAD : The standard format of preparing research articles is APA – American Psychological Association.

Following contents are always present when we start to make research paper –

Abstract – The Standard word limit of abstract according to APA is = 75 – 120 words.

Definition – It is a brief comprehensive summary of Research paper which allows the reader to survey the contents of paper quickly.

Nature or Structure –

Abstract must be -

1. Accurate –which correctly reflects the purpose and content of manuscript
2. Self-contained –
3. Concise and specific – to make sentence informative and as possible as brief.
4. Non-evaluative – to report information objectively.

Introduction

(A) Introduction of Your Research Problem –

It presents specific problem under study.

It describes the research strategy.

(B) Develop the Background –

In this discuss the previous Research completed on this subject.

But in most cases we don't give complete presentation of it.

Assume that an audience is already has general knowledge of field of study.

(A) State the Purpose –

State the Hypothesis.

Explain the purpose behind the Hypothesis.

Method : In this describe the research and include all details on how the study was conducted.

Then identify the sub-sections –

(1) Participants

(2) Measures

(3) Procedure

Each section should be labeled appropriately and placed in BOLD face at left side of page.

(1) Participants – It Describe Number and demographics of participants.

Also be sure to identify source of subject population.

(2) Measures – It explains tests or surveys used for the assessments.

It measures the source of each used measure and its description completely.

(3) Procedure – It gives the details on the way of assessment was conducted.

Be very clear and concise, so that another Researcher would know exactly what to do in order to replicate the study and obtain the similar result.

Results : In this it objectively inform the reader of data collected and the statistical treatment of them. It includes all the tables or figures for further describing the collected data.

In this do not include an evolution or analysis of data.

Discussion : In this discuss the results of experiment.

Analyze data and interpret the implications of data with respect to original Hypothesis

Then compare the results of current study to work of previous research that was discussed sin introduction.

And recommend what should be done next in regard to future research.

References : Generally the references pages will list only source that actually cited in paper.

Appendices : An appendix is the place to put a copy of research instrument.

Signed consent forms or sign-up sheets or the statistical calculations are added.

QUESTION

1. Write short note on-
   (a) Evidence Based Medicines (EBM)
   (b) Research article structuring (IMRAD)

• • •

Chapter 9

RESEARCH DATA PORTALS

CHAPTER OVERVIEW

- Data Mining
- Data Mining Technologies
- Data Mining Techniques
- Research Data Portals -
  1. Dhara (Digital Helpline For Ayurveda Research Articles)
  2. Pubmed
  3. Ayush Portal
  4. Bio-Informatics

(1) DATA MINING

What is Data Mining? Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

What Can Data Mining Do? Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

Market segmentation - Identify the common characteristics of customers who buy the same products from your company.

Customer churn - Predict which customers are likely to leave your company and go to a competitor.

Fraud detection - Identify which transactions are most likely to be fraudulent.

Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.

Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.

Market basket analysis - Understand what products or services are commonly purchased together; e.g., beer and diapers.

Trend analysis - Reveal the difference between a typical customer this month and last.

How Data Mining Works : How is data mining able to tell you important things that you didn't know or what is going to happen next? That technique that is used to perform these feats is called modeling. Modeling is simply the act of building a model (a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers aren't known. Modeling techniques have been around for centuries, of course, but it is only recently that data storage and communication capabilities required to collect and store huge amounts of data, and the computational power to automate modeling techniques to work directly on the data, have been available.

As a simple example of building a model, consider the director of marketing for a telecommunications company. He would like to focus his marketing and sales efforts on segments of the population most likely to become big users of long distance services. He knows a lot about his customers, but it is impossible to discern the common characteristics of his best customers because there are so many variables. From his existing

database of customers, which contains information such as age, sex, credit history, income, zip code, occupation, etc., he can use data mining tools, such as neural networks, to identify the characteristics of those customers who make lots of long distance calls. For instance, he might learn that his best customers are unmarried females between the age of 34 and 42 who make in excess of $60,000 per year. This, then, is his model for high value customers, and he would budget his marketing efforts to accordingly.

(2) Data Mining Technologies

The analytical techniques used in data mining are often well-known mathematical algorithms and techniques. What is new is the application of those techniques to general business problems made possible by the increased availability of data and inexpensive storage and processing power. Also, the use of graphical interfaces has led to tools becoming available that business experts can easily use.

Some of the tools used for data mining are:

Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees—Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Rule induction—The extraction of useful if-then rules from data based on statistical significance.

Genetic algorithms—Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.

Nearest neighbor - A classification technique that classifies each record based on the records most similar to it in an historical database.

(3) Data Mining Techniques

Definition—It is the analytic step of the Knowledge discovery in the database process.

It is interdisciplinary subfield of computer science. It is computational process of discovering patterns in large data sets.

The overall goal of data mining process is – to extract information from data set and transform it into understandable structure for further use.

Actually it is practical machine- which is used for learning tools and techniques with Java and the term data mining was only added for marketing reasons.

The more appropriate term for data mining is data analysis, machine learning, or Artificial Intelligence.

Data mining uses information from past data to analyze the outcome of particular Probability or situation that may arise.

The Data mining works to analyze the data workhouses that are used to store that data. The Ultimate goal is "Prediction."

Predictive data mining is most common type of data mining

It Consist 3 Stages –

(1) Exploration

(2) Model Building and validation

(3) Deployment

(1) Exploration : This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and in case of data sets with large numbers of variables performing some preliminary features selection operations to bring the number of variables to a manageable range.

Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straight forward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity.

(2) Model Building and validation : This stage involves considering various models and choosing the best one based on their predictive performance.

This may sound like a simple operation, but in fact, it sometimes involves a very elaborative process.

The techniques which are often considered the core of predictive data mining include –

Bagging (Voting, Averaging)

Boosting

Stacking (stacked generalizations) and

Meta-learning

(3) Deployment : That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

(4) Research Dataportals

(a) Dhara (Digital Helpline For Ayurveda Research Articles) : DHARA is an online index of articles on Ayurveda published in research journals worldwide.

DHARA (Digital Helpline for Ayurveda Research Articles) is the first comprehensive online indexing service for research articles published in the field of Ayurveda.

The DHARA project is being implemented by AVT Institute for Advanced Research, the research wing of the Ayurvedic Trust, AVP Group of Institutions, Coimbatore, Tamil Nadu, India.

The Ayurvedic Trust was established in 1950. It is the philanthropic wing of AVP Group of Institutions. The Trust has been promoting Ayurvedic Research right from its inception and the Hospital associated with it has evolved into a referral center for Rheumatology in Ayurveda. In 1976, Trust had signed an agreement with World Health Organization (WHO) and Indian Council for Medical Research for Clinical Research in Rheumatoid Arthritis. In 2003, National Institutes of Health (NIH), USA awarded PICRC Research Grant to the Ayurvedic Trust through the University of Washington to study Ayurveda in Rheumatoid Arthritis. Both these initiatives were first of its kind activities in the field of Ayurveda.

The Ayurvedic Trust has now received the Center of Excellence grant from the Department of AYUSH, Ministry of Health and Family Welfare, Govt. of India as a Clinical Research Hospital specializing in Rheumatology.

The first phase of the DHARA project was funded by Central Council for Research in Ayurvedic Sciences, Department of AYUSH, Ministry of Health and Family Welfare, Government of India from August 2010 to July 2011.

Introduction : DHARA is a free indexing service for papers published in research journals in the field of Ayurveda. There is no fee for journals to index published papers. And there is no fee either for users to access the DHARA database.

DHARA does not make all research papers on Ayurveda available free of cost. Free access to full text of the paper depends on journal policy. Journals that advocate an open access policy allow free access to full text. Other journals restrict access to full text by charging a fee. DHARA only indexes the titles or abstracts or both, depending on journal policy. In majority of the cases, both titles and abstracts can be accessed. DHARA provides a link to journal website pointing to the full paper and does not directly store the full text of the published papers.

Even though DHARA does not always directly provide access to full papers, it serves as a powerful search tool to search for prior research work on any topic in a well-structured and systematic way. It helps to track research work in Ayurveda in a comprehensive manner. The controlled vocabulary search can generate precise information on different parameters of research in a particular field of interest. Moreover, the link to the full text helps to locate the article precisely.

DHARA covers only research journals that publish papers after a process on internal and external review. The journal must comply with the basic guidelines of the International Committee of Medical Journal Editors to become eligible to be indexed in DHARA

DHARA does not index PhD or PG theses or dissertations. Only full papers published in research journals are included for indexing.

Features : DHARA mimics PubMed in some respects and this is intentionally done for the sake of convenience of users who are familiar with PubMed.

The main features of DHARA include the

1. Wild Keyword Search Option,
2. Advanced Search with Boolean Operators,
3. Search Field Tags,
4. Limiting Search with Controlled Vocabulary,
5. Journals Area for listing journals by keyword search or in alphabetical order to access basic information on journals,
6. Authors area for listing authors by keyword search or in alphabetical order to search articles by author names,
7. Building URLS that can be stored for future use,
8. Options for sorting search results and

9. Exporting the search results as text file.

(2) PubMed

Definition : It is free online search engine for accessing the Medline database of citations & abstracts of biomedical search articles.

Authority : PubMed is a database developed by

NCBI (National Center for Biotechnology Information) &

NLM (National Library of Medicine)

It was developed as a part of EIRS.

Content : It has over 18 million of citations from Medline & other life science journals for biomedical articles along with free of charge to access Medline.

Fields : It covers data of fields related to medicines, nursing & other health disciplines. Also it fully covers bio-medical sciences like bio-chemistry & cell biology.

Scope : It also offers –

i. Old citations –

ii. *In process citations- those being reviewed for inclusion in Medline & if in scope, subsequently are indexed with Mesh vocabulary. It carries status tag [PubMed- In Process]*

iii. *Publisher Supplied citations – that are supplied by publisher directly electronically to PubMed. It carry tag [PubMed- as supplied by publisher]*

iv. *Some Life-science Journals – that submits full text to PubMed center and may not have been recommended for inclusion in Medline.*

Searching Process

(a) Searching article or whole data –

i. *It is comprehensive search.*

ii. *It is quick process of searching.*

iii. *In this process clinical queries or systematic reviews are searched. It is designed for clinicians & has built in "search filter."*

iv. *Also the Related articles are searched.*

v. *Mapping to MeSH& sub headings.*

vi. Searching with tags & Boolean operations will be done.

(b) Journal database can be searched by –

i. Subject

ii. Journal title

iii. Title abbreviation

iv. NLM identity

v. ISO abbreviation

vi. International Standard Serial Number.

Tags

[au] – author. Ex- Miller R.A. [au]

[dp] – date published

[ip] – issue, part or supplement

[in process]

[published] etc.

Boolean operators

AND (intersection)

OR (union)

NOT (exclusion)

[NOT – may generate false negative result.]

(3) AYUSH Portal : The AYUSH Research Portal is meant for disseminating the information related to the AYUSH and the current research updates purely meant for academic purpose. Users can search AYUSH terminology, research articles, journals etc. You can also search information the Ayurveda, yoga, naturopathy, Unani, Siddha, Homeopathy etc. Information about the clinical research, pre-clinical research, drug research and fundamental research.

AYUSH Research portal is meant for dissemination of Research findings in the domain of Ayurveda, Siddha, Unani, Yoga & Naturopathy and Homeopathy researchers and allied faculties.

Main aim of the portal is to show case the research findings in an organized fashion and preempt duplication of work to encourage interdisciplinary research and generate evidence for wider acceptance of these systems worldwide.

Portal provides information under the headings

1. Standard Treatment Guidelines
2. Preventive primitive health
3. Preclinical and Clinical Studies
4. Literary and Fundamental Research
5. Drug standardization
6. Local health traditions
7. Drug monographs
8. Formulary of India and other formularies
9. References from Classical text books and
10. Plant monographs.

The information provided is categorized as per individual AYUSH medical systems against standard set of Medical conditions, based on WHO disease classification ICD-10, and navigation is provided according to ICPC's (International Classification of Primary Care) in 17 disease categories.

The Portal provides the relevant indigenous name of the medical conditions pertaining to each medical system, which further adds to clarity. The users can browse from a dashboard which provides access to all the categories with visually clear buttons meant for the purpose. On selection of the displayed results, the user will get information in the form of an abstract, and where ever possible, full text of the article.

An important value addition of the portal is categorization of Clinical Research work into Grade A, B, C based on "General Guidelines for methodologies on Research and Evaluation or Traditional medicine published by World Health Organization (WHO). The grading allows the researchers and also policy makers to assess the current R&D status of AYUSH and plan further to fill in the gaps.

The portal content development is a three tier exercise involving rigorous collaborative content uploading and editing process to ensure high quality content. The application has been developed by Central Council for Research in Ayurveda and Siddha (CCRAS), and the content is being provided by all AYUSH Research councils, Drug standardization laboratories, and National Institutes.

AYUSH Research Portal accessible through: http://ayush.ap.nic.in/

In the wake of stiff competition at home as well as worldwide for a legitimate recognition for AYUSH systems, the portal will serve as platform to further the cause of Strengthening these systems.

(4) Bio-Informatics

Definition : Study of information process in biotic system is called as "Bio-informatics."

Introduction : In 1978 Pauline Hogeweg coined a term – Bio-informatics firstly.

"David Lipmann" (Director of National Centre for Biotech information) – is a Father and Mother of Bio-informatics.

The Primary Goal of Bio-informatics is to understanding of biological processes.

The Common activities include mapping and analyzing of DNA and protein sequence.

'Rosalind' is educational resource and web project for learning the bio-informatics through problem solving and computer programming.

The Basic bio-informatics services are classified by EBI into 3 stages–

(1) Sequence Search Services

(2) Multi Sequence Alignment

(3) Biological Sequence Analysis

A broad sub category under bio-informatics is Structural bio-informatics.

Bioinformatics, the latest branch of biological science, is the science of using information technology to understand biology. In a literal sense Bioinformatics is the integration of life sciences and information science. It is the science of creating and managing biological databases to keep track of and eventually simulate the complexity of living organisms. Thus, biology, computer science and information technology merge into a single discipline in Bioinformatics. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein

sequences to compare them and creating and viewing 3-D models of protein structures.

Scope

(1) It became an important part of many areas of biology like experimental biology, molecular biology, bio-informatics techniques etc.

(2) In field of genetic or genomics, their sequencing annotating genomes and mutation.

(3) Analysis of gene and protein expression and regulation.

(4) Comparing Genetic and genomic data.

(5) Used in many areas of computer science, mathematics and engineering to process a biological data.

The Major activity in Bio-informatics is to develop a software tools to generate useful biological knowledge.

CENTRES

Main Objectives

Being located in a premier institution in the country, the center plays a significant role in disseminating much of the information required for various aspects of research in life sciences locally and in the whole country.

Some of the main objectives of the Centre are:

- To coordinate collection, retrieval, collation, distribution, delineation & deciphering of biological related information with particular emphasis to molecular biology.
- To carry out original research work in the broad areas of sequence and structural bioinformatics.
- To create and maintain specialized derived databases in molecular biology.
- To develop bioinformatics software tools with an aim to provide wide access to the research community in the country.
- To provide local access to various databases on biomolecules in the fields of biotechnology, molecular biology and genetic engineering.
- To make available the latest software to users in the new areas of sequence and structural bioinformatics.
- To conduct training programs to make users aware of the available facilities and their applications.

- To provide guidance to researchers in using the above databases and software.
- To carry out developmental work in response to requests from users.
- To establish good linkages among scientists within the institute and with those working outside the organization through networking and by providing broad bandwidth Internet accessibility.
- To build up information resources, prepare databases pertaining to spices and to develop relevant information handling tools and techniques.
- To continuously assess information requirements, organize creation of necessary infrastructure and to provide information and computer support services to the scientists.
- To evolve and implement programs on education of users and training of scientists responsible for handling of biological information.

The Bioinformatics activity in India was initiated in 1987 with the establishment of a National Bioinformatics Network known as Biotechnology Information System of India (BTIS) by the Department of Biotechnology (DBT), Government of India. It is a distributed information network comprising of 10 Distributed Information Centres (DICs) and 50 Distributed Information Sub-Centres (DISCs).

Centres of Excellence : BTISNet consists of Seven Centres of Excellence (COE) in Bioinformatics. The centres are:

Apex Bioinformatics Centre, DBT, New Delhi

University of Pune, Pune

Jawaharlal Nehru University (JNU), New Delhi

Madurai Kamaraj University (MKU), Madurai

Indian Institute of Science (IISc), Bangalore

Bose Institute, Kolkatta

Super Computing Facility (IIT), New Delhi

Distributed Information Centers (DICs) : Eleven DICs have been established with the task of providing discipline-oriented information to all institutions belonging to the branch as well as other institutions and individual users interested in particular subject related to Biotechnology. These are listed below:

List of Distributed Information Centers (DICs)

Anna University
Centre for Cellular & Molecular Biology
Indian Agricultural Research Institute
Institute of Microbial Technology
Kerala Agriculture University
M. S. University of Baroda
National Brain Research Centre
National Institute of Immunology
North Eastern Hill University, Shillong
Pondicherry University
University of Calcutta

QUESTION

1. Importance of IT in data mining
2. Write short note on Data mining technique
3. Write short note on-
   (a) DHARA
   (b) PubMed
   (c) AYUSH Research Portal
   (d) Bio Informatics

• • •

chapter 1

INTRODUCTION OF MEDICAL STATISTICS

CHAPTER OVERVIEW

- Introduction of Statistics
- Definitions of Statistics
- Scope of Statistics
- Importance of Statistics

Introduction : When we hear the word 'statistics', we visualize numerical facts presented in table format. Statistics as a subject doesn't have this limited perspective.

The word **'Statistics'** derived from the Latin word '*Status*' or German work '*Statistik*', which means 'Political state'.

Statistics is a branch of mathematics, which has great practical value as tool in experimental research. Statistical methods are essentially ways of handling information obtained by repeated measurements.

Statistics is one of the devices by which men try to understand the generality of life. It is a body of mathematical techniques or processes for gathering, organizing, analyzing, and interpreting numerical data.

Statistics is a basic tool of measurement, evaluation and research.

Sir Ronald A. Fisher, known as the **'Father of Statistics'** placed statistics on a very sound footing by applying it to various diversified fields, such as genetics, biometry, education, agriculture etc.

Definition : Many authors based on its widened utility defined in many ways i.e. statistical data and statistical methods etc.

i. "Statistics is the science of estimates and probabilities"
ii. "The science, which deals with the collection, analysis and interpretation of numerical data"
iii. "Science of counting (or) science of average"
iv. "Controlled, objective methods by which group trends are abstracted from observations on many separate individuals are called statistical methods"

v. "Statistical data may be considered broadly as observations made on subjects."

vi. Bowley defined statistics as "Numerical statements of facts in any department of enquiry placed in relation to each other."

Steps in Statistics : "Statistics is the study of collection, tabulation, analysis, interpretation & presentation of data."

Data is nothing but the information which we receive in either experimental study, observational study or in retrospective study like records.

Collection of data comprises of design of experiment & different sampling methods, which ensure the validity of representativeness and quality of data. For the case of understanding, data is represented in different table formats. This process is known as Tabulation.

Data is then examined using different statistical methods. This process is known as Analysis.

Data analysis can be used for understanding the nature of data, extraction of information, its prediction and quality control. The first step in the data analysis is formulating the problem in mathematical form .

Ex-

i. While conducting the ayurvedic experiments, many times patients discontinue their treatment leading to incomplete data. In this case, we can use missing plot technique to estimate these missing observations, saving the need for repetitions of an experiment.

ii. Optimum dose of drug can be predicted by noting down the response at various amounts of drugs & modeling using regression analysis. Such prediction is useful in finding the tolerance level of drug which can be harmful to the human being or an animal on which the actual experiment is carried out.

iii. Basti treatment is helpful in gaining or losing the weight. We can prove the effectiveness of basti treatment by recording weights of patients before and after treatment by using t-test.

Interpretation of the result obtained from analysis means to decide the meaning of results i.e. what the data says about our experiment.

The last but most important step is :

Presentation of the data or result. This is a part of analysis which is presented in most of the research papers.

BIO-STATISTICS

Definition : it is a branch of statistics which related to the study of collection, tabulation, analysis, interpretation & presentation of biological data.

This term Bio-statistics is used when the statistical tools are applied to the data which is related to biological science like medicine.

Medical Statistics : "The utility and application of statistical methods in the field of medical sciencefor its wider acceptance is called as "Medical Statistics."

"It is the science of summarizing, collecting, presenting and interpreting data in medical practice and using them to estimate the magnitude of associations and test hypotheses.

It has a central role in medical investigations. It not only provides a way of organizing information on a wider and more formal basis than relying on the exchange of anecdotes and personal experience, but also takes into account the intrinsic variation inherent in most biological processes.

Medical statistics deals with applications of statistics to medicine and the health sciences, including epidemiology, public health, forensic medicine, and clinical research.

Medical statistics go under different names when they applied in different field's like-Health statistics, Medical statistics, vital statistics etc.

Scope of statistics :

To a very striking extent our culture has become a statistical culture. Even on the most elementary level it is impossible to understand psychology, sociology, economics, finance, or the physical sciences without some general idea of the meaning of an average, of variation, of relationship, of sampling, of how to read charts and tables.

It is now finding wide applications in almost all sciences – social as well as physical, such as biology, psychology, education, economics, business management, astronomy, medical science etc.

Statistical information concerning population changes are essential to the long-range planning of such agencies as housing and traffic authorities, public utilizes, and school boards.

In medical field some part of statistics must be learned to understand their application in the diagnosis, prognosis& treatment of the diseases.

Importance of statistics in research :

i. Statistics is the science of collection, analysis, interpretation or explanation, and presentation of data. It has wide usage in the field of research. In fact all the data collection and interpretation techniques used in Research are part of statistics.

ii. It makes use of descriptive statistics for collection of data and inferential statistics for drawing inferences from this set of data.

iii. The subject called research statistics & statistics is very important in research because that is the backbone of your research.

iv. The Numbers gives an easy idea of how you conducted your research.

v. Statistics provides a platform for research as to; How to go about your research, either to consider a sample or the whole population, the Techniques to use in data collection and observation, how to go about the data description (using measure of central tendency).

vi. To wrap it up, statistics as a science of data collection, analysis, interpretation, explanation and presentation will guide you in research for proper characterization, summarization, presentation and interpretation of your research result for proper action.

IMPORTANCE OF MEDICAL STATISTICS

(i) In Anatomy & Physiology : *Medical statistics plays important role for the declaration of an individual as a Healthy or normal individual by various methods.*

Ex – Pulse rate – mean pulse rate is 72 but the limits of normality on either side is calculated by using the various statistical methods& then only we declare that individual as Healthy. The same techniques can be used for the other variables like weight, height of an individual, height weight ratio, its BMI etc.

(ii) In Pharmacology : *To calculate dose of a drug.*

To compare actions of two different drugs.

(iii) In Medicine : *To compare the efficacy of new drug by comparing it with standard drug.*

To find an association between two things like relation between smoking and cancer.

(iv) In Public Health : To find the efficacy of a vaccine in the specific area.

To find out the actual cause of an epidemic diseases.

IMPORTANCE OF MEDICAL STATISTICS IN AYURVEDIC RESEARCH

(i) In Sharir Kriya : According to ayurveda, each person is said to have different types of prakruti& accordingly these are physiological variation in them. These variations in prakruti can be studied by application of statistics.

Ex – the skin of vataprakruti person is Ruksha (rough) as compared to kaphaprakruti. So thus it helps to find out whether the nature of skin of that person is due to chance or due to natural variations.

(ii) In Sharir Rachana : Statistics can be applied to define the normal dimensions of various organs, number of structures like bones, arteries, marmas etc.

(iii) In Rog-nidan : When there is confusion about the diagnosis of disease, the physician has to identify the specific symptoms of disease i.e. pratyatmalakshana

Ex – evening rise of body temperature is found in malaria, Koplik's spots were found in measles etc. found in the patients was due to chance or was present in all cases.

This proportional incidence of symptoms indicates whether it is a characteristic of that disease or not.

(iv) In Agadtantra : It helps to estimate the action of vishadravya in vitro & in vivo and to decide it's optimal dose.

(v) In Rasashastra : The difference in the melting point, boiling point, specific gravity etc. of the compound can be studied statistically and used in their identification.

Statistical tests can be applied to the trials conducted on purification, dosage & duration of various rasa kalpas to avoid their toxic effects.

(vi) In Swasthavritta : The occurrence of the epidemic disease in a particular area is by chance or due to some reason can be studied.

(vii) In dravyaguna : To find out the action of drug - a drug is given to

animal or human to see whether the changes produced are due to drug or by a chance.

Also it can be used to compare the action of same drug at different dosage.

(viii) In Kayachikitsa : It is used to compare the efficacy of the line of treatment in a particular disease. The difference in the percentage of patients dies, cured or relieved in the experiments is studied statistically.

(ix) In stri-rog & prasutitantra : Theusefulness of the drugs & various methods mentioned in ayurveda for improving the fertility chances can be studied.

Statistical tests can also be applied to the trials conducted on various oral & other contraceptives for its safety & effectiveness.

(x) In balrog : The efficacy of various immunization vaccines can be studied statistically for prevention of disease in children.

QUESTIONS

1. Define statistics & its scope in ayurvedic research.
2. Write the importance of medical statistics.
2. Explain bio statistics.

• • •

Chapter 2

COMMON STATISTICAL TERMS & NOTATIONS

CHAPTER OVERVIEW

- Population
- Sample
- Data
- Variable
- Normal distribution

(1) Population : A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

In order to make any generalizations about a population, a sample, that is meant to be representative of the population, is often studied. For each population there are many possible samples. A sample statistic gives information about a corresponding population parameter.

For example, the sample mean for a set of data would give information about the overall population mean.

It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included.

Ex – The population for a study of infant health might be all children born in the India in the 1980's. The sample might be all babies born on 7th May in any of the years.

(2) Sample : A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group.

A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling.

Also, before collecting the sample, it is important that the researcher

carefully and completely defines the population, including a description of the members to be included.

Ex – The population for a study of infant health might be all children born in the India in the 1980's. The sample might be all babies born on 7th May in any of the years.

(3) Data : A set of values recorded on one or more observational units.

Or

Data can be defined as a collection of facts or information from which conclusions may be drawn.

Data are measurements or observations that are collected as a source of information. There are a variety of different types of data, and different ways to represent data.

Ex – The number of peoples in India,

Number of calls received by the emergency services each day,

The value of sales of a particular product,

The number of times India has won a cricket matchthese are all examples of data.

i. **Data unit** – is one entity (such as a person or business) in the population being studied, about which data are collected. A data unit is also referred to as a unit record or record.

ii. **Data item** – is a characteristic (or attribute) of a data unit which is measured or counted, such as height, country of birth, or income. A data item is also referred to as a variable because the characteristic may vary between data units, and may vary over time.

An observation is an occurrence of a specific data item that is recorded about a data unit. It may also be referred to as datum, which is the singular form of data.

An observation may be numeric or non-numeric (categorical).

For ex – 173 is a numeric observation of the data item 'height (cm)', whereas 'India' is a non-numeric (categorical) observation of the data item 'country of birth'.

iii. **Data Set** – is a complete collection of all observations.

The following table is an example of a dataset:

	Age (years)	Sex	Income($)
Person 1 (John Smith)	18	m	50000
Person 2 (Joe Bloggs)	46	m	41666
Person 3 (Sally Jones)	20	f	55000
Person 4 (Landa Lee)	22	f	50000
Person 5 (Harry James)	19	m	35000

Data Item

Data Unit Person 2

Numeric observation of the data item 'income'

Non-numeric (categorical) observation of the data item 'sex'

TYPES OF DATA :

(A) Qualitative & Quantitative data.

(B) Primary & secondarydata.

(C) Discrete & Continuesdata.

(D) Grouped & ungroupeddata.

(A) Qualitative & Quantitative data

Data collected about a numeric variable will always be quantitative and data collected about a categorical variable will always be qualitative. Therefore, one can identify the type of data, prior to collection, based on whether the variable is numeric or categorical.

(i) Qualitative data : The Data which collected about a categorical variable will always be qualitative.

Qualitative data is non- numeric in nature.

Qualitative data are often termed categorical data.

Ex – {Poor, Fair, Good, Better, Best}, colors (ignoring any physical causes), and types of material {straw, sticks, bricks} are examples of qualitative data.

(ii) Quantitative data : The Data collected about a numeric variable will always be quantitative.

Quantitativedata isnumeric in nature.

Ex – If data are collected on annual income (1 lakh, 2 lakh, 3 lakh etc.), number of members in family (3, 4, 5, 6 etc.) are examples of quantitativedata.

(B) Primary & secondary data

According to the source of collection, the data can be classified as Primary or secondary data.

(i) Primary Data : The Data which is collected by investigator for first time for doing statistical analysis by him is known as 'Primary Data.' Some examples of primary data are: voters list, data collected in census, data collected from questionnaire etc.

(ii) Secondary Data : The Data which is obtained from sources like journals, Government hospital records etc. It is not directly collected by the analyst.

(C) Discrete & Continuous data

According to the nature of that specific numerical data, it is classified into Discrete or continuous data. In other words we can say that it is a subdivision of quantitative data as it contains only numerical data.

(i) Discrete data : Discrete data is numeric data that have a finite number (countable) of possible values.

Data contains whole numbers and Always Qualitative.

Ex – A classic example of discrete data is a finite subset of the counting numbers, {1, 2, 3, 4, and 5} or Number of patients dying due to cancer.

(ii) Continuous data : Continuous data have infinite possibilities.

The real numbers are continuous with no gaps or interruptions.

Ex – Physically measureable quantities of length, volume, time, mass, etc. are generally considered continuous data.

(D) Grouped & ungrouped data

According to presentation of data it can be divided as Grouped data or ungrouped data.

(i) Grouped Data : The data which is presented in groups named as Grouped Data.

Ex – The Hb % of 10 patients can be represented as

13.2 gm% - 3 patients, 13 gm% - 2 patients, 13.4 gm% - 5 patients etc.

Hb in gm%	Frequency of patients
13.2	3
13	2
13.4	5

(ii) Ungrouped Data : The data which is presented individually (without groups) named as 'ungrouped data.'

Ex – The Hb % of 10 patients can be represented as

13.2 gm%, 13 gm%, 13.2 gm%, 13.4 gm%, 13.4 gm%, 13 gm%, 13.4 gm%, 13.4gm%, 13.2 gm% and 13.4 gm%

No. of the patient	1	2	3	4	5	6	7	8	9	10
Hb in gm%	13.2	13	13.2	13.4	13.4	13	13.4	13.4	13.2	13.4

(4) Variable : A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.

Ex – Age, sex, business income and expenses, country of birth, capital expenditure, class grades, and eye color and vehicle type etc.

It is called a variable because the value may vary between data units in a population, and may change in value over time.

For example; 'income' is a variable that can vary between data units in a population (i.e. the people or businesses being studied may not have the same incomes) and can also vary over time for each data unit (i.e. income can go up or down).

Types : - There are different ways variables can be described according to the ways they can be studied, measured, and presented.

(A) Numeric variables : Numeric variables have values that describe a measurable quantity as a number, like 'how many' or 'how much'. Therefore numeric variables are quantitative variables. Hence the data collected for a numeric variable are quantitative data.

Numeric variables may be further described as -

(1) Continuous

(2) Discrete

(1) Continuous variable : It is a numeric variable.

Observations can take any value between a certain set of real numbers. The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows.

Ex – height, time, age, and temperature

(2) Discrete variable : It is a numeric variable.

Observations can take a value based on a count from a set of distinct whole values. A discrete variable cannot take the value of a fraction between one value and the next closest value.

Ex – the number of registered cars, number of business locations, and number of children in a family, all of which measured as whole units i.e. 1, 2, 3 cars

(B) Categorical variables : Categorical variables have values that describe a 'quality' or 'characteristic' of a data unit, like 'what type' or 'which category'.

Categorical variables fall into mutually exclusive (in one category or in another) and exhaustive (include all possible options) categories. Therefore, categorical variables are qualitative variables and tend to be represented by a non-numeric value. Hence the data collected for a categorical variable are qualitative data.

Categorical variables may be further described as -

(i) Ordinal

(ii) Nominal

(1) Ordinal variable : Ordinal variable is a categorical variable.

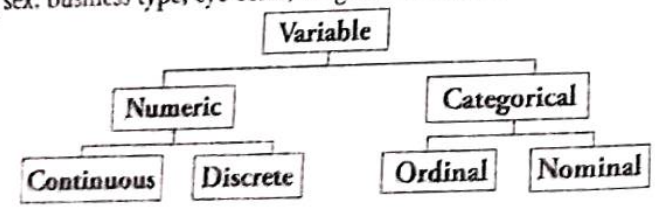
Observations can take a value that can be logically ordered or ranked. The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category.

Ex – academic grades (i.e. A, B, C), clothing size (i.e. small, medium, large, extra-large) and attitudes (i.e. strongly agree, agree, disagree, and strongly disagree).

(2) Nominal variable : Nominal variable is a categorical variable.

Observations can take a value that is not able to be organized in a logical sequence.

Ex – sex, business type, eye color, religion and brand



(5) Normal Distribution: - Notation $N(\mu, \sigma^2)$**:** The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss.

As you will see in the section on the history of the normal distribution, although Gauss played an important role in its history, Abraham de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about "the normal distribution" since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations.

Figure shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3.

These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

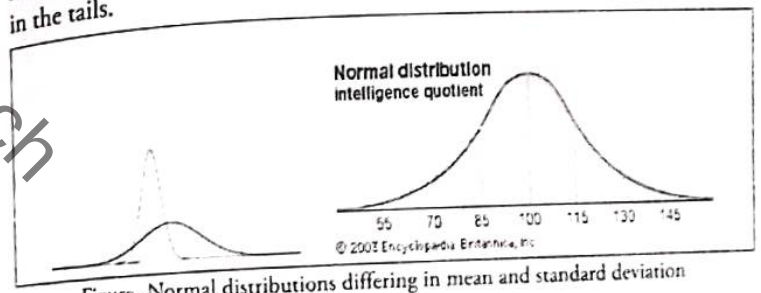


Figure- Normal distributions differing in mean and standard deviation

1. Seven features of normal distributions are listed below.
2. Normal distributions are symmetric around their mean.
3. The mean, median, and mode of a normal distribution are equal.
4. The area under the normal curve is equal to 1.0.
5. Normal distributions are denser in the center and less dense in the tails.
6. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
7. 68% of the area of a normal distribution is within one standard deviation of the mean.
8. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

A normal distribution is:

$$f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter in this definition is the mean or expectation of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is therefore σ^2. A random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.

If $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution or the unit normal distribution denoted by N(0,1) and a random variable with that distribution is a standard normal deviate.

The normal distribution is the only absolutely continuous distribution all of whose cumulants beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a given mean and variance.

Statistical Symbols

	Population - parameter	Sample - statistic
size:	N	n
center:		
mean	μ "mu"	$\bar{X}$ "x bar"
median	n/a	M or $\tilde{X}$ "x tilde"
proportion	π "pi"	p
	(p in text)	(p in text)
Variability		
variance	σ^2 "sigma squared"	s^2 "s squared" = $\Sigma(x - \bar{X})^2/(n-1)$
standard deviation	σ "sigma"	s
range	n/a	n/a
interquartile range	n/a	IQR = $Q_3 - Q_1$
relative standing:		
z score = (x - mean)/SD	Z	z

	Population - parameter	Sample - statistic
the # of SD's from the mean		
for bivariate data:		
correlation coefficient	ρ "rho"	r
	These are fixed numbers; usually unknown.	These vary from sample to sample. We use them to estimate the population parameters.

Standard Normal Distribution Notation: $Z \sim N(0,1^2) \rightarrow Z$, a random variable, is distributed normally with mean, $\mu = 0$, variance, $\sigma^2 = 1^2$, and standard deviation, $\sigma = 1$.

Non-Standard Normal Distribution: $X \sim N(\mu_x,\sigma_x^2) \rightarrow X$, a random variable, is distributed normally with mean, μ, variance, σ^2, and standard deviation, σ.

Sampling Distribution of the Sample Mean from a Normal Population: $\bar{X} \sim N(\mu_x, \sigma_x^2/n) \rightarrow \bar{X}_n$, a random variable calculated from a sample of size n, is distributed normally with mean, $\mu_{\bar{X}} = \mu_x$ (the mean of the parent population), variance, $\sigma_{\bar{X}}^2 = \sigma_x^2/n$ and standard deviation, $\sigma_{\bar{X}} = \sigma_x/\sqrt{n}$.

Sampling Distribution of the Sample Proportion: $p_n \sim N(\pi, \frac{\pi(1-\pi)}{n}) \rightarrow p_n$, a random variable, is distributed normally with mean, $\mu_{p_n} = \pi$, variance, $\sigma_p^2 = \frac{\pi(1-\pi)}{n}$ and standard deviation, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$.

Standard Notation for Statistics

A distinction is made between a statistic that is computed on everyone in a population and the same statistic that is computed on everyone in a sample drawn from the population.

A statistic computed on everyone in the population is called a population parameter.

A statistic computed on everyone in a sample is called a sample statistic.

The population mean is designated by the Greek letter mu, whereas the

sample mean is designated by an X with a bar over the top (read X bar). Both are illustrated below.

Population Mean = μ

Sample Mean = $\bar{X}$

A similar distinction is made for standard deviation, which is a measure of variability. The population standard deviation in indicated by the lower case Greek letter sigma, whereas the sample standard deviation is indicated by the lower case letter s, as shown below.

Population Standard deviation = σ

Sample Standard deviation = s

The lower case letter *r* is used to designate a correlation. If there is any doubt about which two variables were used to compute the correlation, the two variables are listed as subscripts. For example, r_{XY} indicates the correlation of X and Y.

Parameter name	Population Parameter Symbol	Sample Statistic
Number of Cases	N	n
Mean	μ (mu)	$\bar{X}$ (Sample mean)
Proportion	π (Pi)	P (Sample proportion)
Variance	σ (Sigma - square)	s^2 (Sample Variance)
Standard deviation	σ (Sigma)	s (Sample standard deviation)
Correlation	ρ (rho)	r Sample correlation
Regression Coefficient	β (beta)	b (Sample regression coefficient)

QUESTIONS

1. Explain in brief —
   (a) Population (b) Sample
   (c) Data (d) Variable
   (e) Normal distribution
2. Explain population with its notations.
3. Explain normal distribution.

• • •

Chapter 3

COLLECTION & PRESENTATION OF DATA

CHAPTER OVERVIEW
- Collection of Data
- Classification of Data
- Presentation of Data
- Tabular
- Drawing–Graphical & Diagrammatical

(A) COLLECTION

Data collection is collection of data from their sources.

The main sources for collection of data in medical statistics are –

(a) Experiments

(b) surveys

(c) records

(a) Experiments : Medical experiments performed in lab of Physiology, Biochemistry, Pharmacology and clinical pathology or in hospital wards for investigation and fundamental.

Utility - (a) Dissertation

(b) Thesis

(c) Scientific papers of publication I scientific journals and books.

(b) Survey – It is carried out for epidemiological studies in the field and also used in operational Research.

Utility - It provide useful information on

Changing trends in health status, morbidity, mortality, nutritional status, investigation Hazards etc.

Provide feedback which may be expected to modify policy and system itself and lead to redefinition objects.

Provide trimly warning of public health Hazards.

(c) Records – They are maintained as routine in registers / books over long period of time for various purposes.

Ex – Vital statistics – of Birth
Marriage
Deaths
Illness in hospitals

(B) CLASSIFICATION

(I) **According to Source** - Data may be classified according to source
(a) Primary data
(b) Secondary data.

(a) Primary – The Data which is collected by investigator for first time for doing statistical analysis by him is known as 'Primary Data.'

Merit –it is self-collected, so quality is ensured.

Demerit - Time consuming and costly.

(b) Secondary – The Data which is obtained from sources like journals Government hospital records etc.

It is not directly collected by the analyst.

Merit - It is Time and cost friendly.

Demerit - Some bias may exist.

(II) Qualitative and Quantitative Data -

(a) Qualitative data –

Definition – The Data which can't be measured in numbers but represents particular characteristics is called as 'Qualitative Data.'

Ex. – Sex - Male / Female
Blood group - A, B, AB, O

It is classified by counting individuals having some characteristics of attribute and not measurement.

There is only one variable.

Data is discrete in nature.

The result obtained is expressed in ratio or proportion or % or rate.

In qualitative date characteristics do not vary with vary in frequency.

Such data is mostly collected in

(a) Pharmacology – to find action of drug

(b) Clinical practice – to test or compare efficacy of drug.

Statistical methods which are used to analyze such data are-
SE of proportion
X^2 test.

(b) Quantitative data –

Definition – The Data which can be expressed in numeric forms is called as 'Quantitative data.'

It has magnitude.

It is obtained from characteristic variable which is also called as "Continues Data"

The characteristics are measured either on interval or on ratio scale.

Such statistics are mostly collected in Anatomy and Physiology.

Ex –
- In health – to define normal.
- To find limits of deviation from normal in healthy person.

Some of the statistical methods employed in analysis of such data are –
Mean
Range
Standard Deviation
Coefficient of Variation
Correlation coefficient

(III) Grouped and Ungrouped –

(1) **Grouped** – Data is presented in Groups.

Ex. – Hb gm% of 5 men = 13.2 gm% (3 men), 13.9 (2 men)

(2) **Ungrouped** – Data is presented dividedly.

Ex. – above same data
12.3 gm%, 12.5gm%, 12.9 gm%, 13.2 gm% and 13.5 gm%

(IV) Discrete and Continues –

Discrete - Data contains whole numbers and Always Qualitative.
Never comes in fractions.

Ex. – No. of patients dying due to cancer.

Continues - Possibility of getting fractions like 1, 2, 3.4, 3.6 etc.

Ex. – Depending upon requirement wt. can be expressed in decimals.

SOME IMPORTANT TERMS IN FREQUENCY DISTRIBUTION

(1) Class or Class interval : It is a set of values represented by two closing values, known as upper class Limit (UCL) & lower class limit (LCL).

Ex – age groups like 41-50 years, 51-60 years are Classes having LCL = 41, 51 & UCL = 50, 60 respectively.

(2) Class width : It is the difference between upper class limit (UCL) and lower class limit (LCL).

It is denoted by symbol 'h'.

Ex – age groups like 40-50 years, the class width h = 50-40 = 10 i.e. 1 =10

(3) Mid value or Class marks : It is a representative value of a class and can be calculated as making mean of upper class limit (UCL) and lower class limit (LCL).

$$\text{Mid Value} = \frac{UCL + LCL}{2}$$

(4) Frequency :

It is a number of data points belong to each class.

It is denoted by symbol 'f'

Total frequency is denoted by symbol 'N'.

(5) Class boundaries : When the grouped discrete frequency distribution is converted into continuous frequency distribution, we have to add $\frac{1}{2}d$ to UCL & subtract $\frac{1}{2}d$ from LCL. Then the new class limits which are formed called as Class Boundaries, where'd' is the gap between classes.

$$\text{Upper Class Boundary} = UCL + \frac{1}{2}d$$

$$\text{Lower Class Boundary} = LCL - \frac{1}{2}d$$

(6) Cumulative Frequency : We sometimes need the number of data points having 'Less than' or 'More than' the values of particular value. then that time 'Cumulative frequency' helps us.

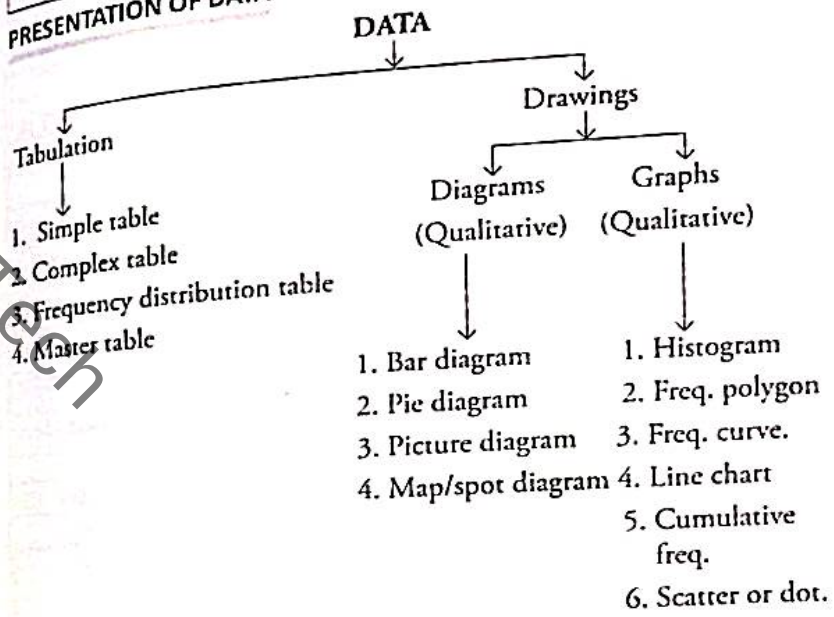
(a) Less than Cumulative frequency (LCF) : It is obtained by adding all the frequencies of **Previous Classes** to the frequency of that class.

(b) More than Cumulative frequency (MCF) : It is obtained by adding all the frequency of **Next Class** to the frequency of that class.

Ex:–

Weight in kg	Tally Marks	Frequency (f)	LCF	MCF
40-50	III	3	3	15+3=18
50-60	⫼ I	6	3+6=9	9+6=15
60-70	⫼	5	9+5=14	4+5=9
70-80	I	1	14+1=15	3+1=4
80-90	III	3	15+3=18	3
Total		18		

PRESENTATION OF DATA



(1) Tabulation : It is a Process of arranging the data in table format.

It gives one shot view of data.

It is first step in presentation and analysis of data.

There is no hard and fast rule for creating a table.

One should have clear idea about facts to be presented and relation between them for proper designing of table.

Types - i. Simple table.
ii. Complex table.

iii. Frequency distribution table.

iv. Master table.

(i) Simple table – It is Very straight forward.

It includes one or two attributes (Characteristics)

It is relatively easy to construct.

The characteristics under observations are fixed.

Number or the frequency of events is small.

Ex. – Location wise distribution of admitted persons according to heart disease.

Location	Heart Disease		Total
	Yes	No	
Rural	40	80	120
Urban	50	30	80
Total	90	110	200

(ii) Complex table : If more than 2 attributes comes for presentation then table become manifold such table is called as Complex table.

Ex. – In previous example – If 3rd attribute "Sex" added.

Location	Heart Disease				Total
	Yes		No		
	Male	Female	Male	Female	
Rural	10	30	40	40	120
Urban	30	20	10	20	80
Total	40	50	50	60	200

(iii) Frequency distribution table : It is most important table.

Large or ungrouped data is presented in small, manageable number.

Types

(a) Discrete frequency distribution (grouped or ungrouped).

(b) Continues frequency distribution (grouped).

(a) Discrete frequency distribution -

(A) Ex. – Age on last b' day –

21, 22, 20, 24, 20, 20, 22, 26, 24, 26, 21, 21, 25, 25, 25, 20, 25, 25, 21, 25

Un-grouped

Age	Tally marks	Freq.
20	IIII	4
21	IIII	4
22	II	2
23	--	0
24	II	2
25	IIII I	6
26	II	2
Total	N	20

Grouped

Age	Tally marks	Freq.
20-21	IIIIII	8
22-23	II	2
24-25	IIIIII	8
26-27	II	2
Total	N	20

(B) Ex- Congenital Malformations: Frequency and Sex Distribution

	Total Cases	Malformed patients	Percentage
Total Births	9386	179	1.90
Still Births	192	9	4.68
Live Births	9194	170	1.84
Male	4813	121	2.51
Female	4570	58	1.26
Ambiguous	3	0	0

(C) Distribution of blood donors with transfusion transmitted infections according to sex

Sex	Hepatitis B		Hepatitis C		Syphilis		HIV		Total	
	No. of patients	%	No. of patients	%	No. of patients	%	No. of patients	%	No. of patients	%
Male	61	58.09	37	35.23	04	3.80	03	2.85	105	92.92
Female	01	12.50	05	62.50	00	00	02	25.0	08	7.08
Total	62	54.86	42	37.16	04	3.53	05	4.42	113	100.0

* Data were collected for a period of 2 years from January 2009 to December 2010.

(b) Continues frequency distribution :

(i) Ex- weight of Patients: - 72, 62.5, 48, 80, 48.2, 61.2, 58.5, 58, 59.5 55.3, 45.8, 63.3, 52, 58.8, 65.8, 81.2, 86, 69.2

Wt. in kg	Tally	Frequency (f)
40-50	III	3
50-60	𝍸 I	6
60-70	𝍸	5
70-80	I	1
80-90	III	3
	Total	18

(ii) Ex – Draw the frequency distribution table for the given ungrouped data. Consider the following ungrouped marks (out of 50) given to 30 students:

30	36	35	42	40	26	23	14	34	40
35	28	33	29	21	47	12	36	34	19
36	32	49	45	20	16	44	30	36	24

Solution :

Frequency distribution table for the given data:

Class	Tally Marks	Frequency
11-15	II	2
16-20	III	3
21-25	III	3
26-30	IIII	5
31-35	𝍸	6
36-40	𝍸	6
41-45	III	3
46-50	II	2
Total		30

The range for the above ungrouped data is 49 - 12 = 37. Normally it is desirable to divide the range into 6 to 10 classes. Consider the class 11-15. If a student scores 11 marks or 15 marks, he will be put in this class.

For this class, 11 is the lower limit and 15 is the upper limit and both are included in the class.

(iii) Ex –weight of persons in age group 25-40 years

Weight (in kg)	No. of males	%	No. of females	%
45-50	6	6.3	25	45.5
51-55	18	19	12	21.9
56-60	46	48.4	8	14.5
61-65	17	17.9	8	14.5
65-70	6	6.3	2	3.6
71-75	2	2.1	0	0
Total	**95**	**100**	**55**	**100**

Rules for Frequency distribution table :

1. Class interval between the groups should not be too broad or too narrow.
2. The number of classes should not be too many or too few.
3. The class interval should be same throughout the observations.
4. The headings must be clear. Ex- Age in years or in months or Height in Feet or in cm etc
5. Groups shouldbe tabulated in ascending or descending order.

(iv) Master Table : It is full fledge table giving detail of each experimental unit one below other.

All column names are either variable or attributes and each row represents a "record".

Pt. No.	Name	Age	Ht.	Wt.
1.	A.V. Patil	28	162	52
2.	S.S. Patil	32	164	48
3.	S.S. Mote	48	168	68
4.	S.D. Kore	25	163	55
5.	J.K. Jadhav	22	159	56

Rules for Table presentation

1. Table must be numbered.
2. Brief and self-explanatory title.

3. Heading of columns and rows must be clear.
4. Table should not be too large.
5. Comparative figures should be placed.
6. Foot notes should be given.
7. Classes should be clearly defined.
8. Non – overlapping.

(2) DRAWINGS : Pictures are easier to understand than numbers therefore charts and graphs prove to be very useful presentation tool.

Graphs are plotted on a special paper known as graph paper. They are generally used with quantitative data and are more precise than diagrams.

Diagrams on the other hand are useful for visual presentation of qualitative data. Diagrams are used for comparative study.

(I) Diagrams : Sometimes our goal is to present just summary of data in such situation diagrams very useful. They generally used with qualitative data.

(a) Bar diagrams : It is useful for presentation of qualitative data.

It is easy to prepare and it used for comparing the categories of mutually exclusive discrete data.

A Length of bars which is drawn vertically or horizontally indicates the frequency.

Bars may be drawn in ascending or descending order of magnitude or in serial order of events.

It consists of group of rectangles or bars which are properly spaced from one another.

There are 3 types of bar diagrams –

(i) Simple bar diagram
(ii) Multiple bar diagram
(iii) Proportional bar diagram

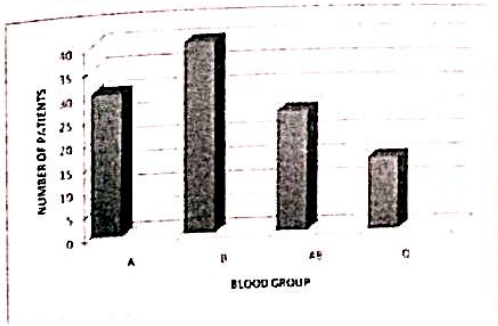
(i) Simple bar diagram : It is simplest form.

It is used for comparative study of two more values of a 'Single Attribute'.

Bars can be arranged vertically or horizontally.

Problem : Represent the following data in simple bar diagram.

Blood group	No. of patients
A	30
B	40
AB	25
O	15

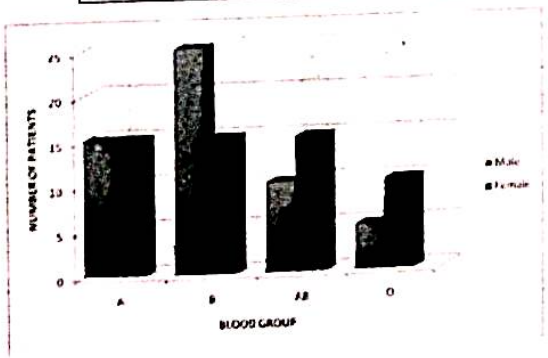


(ii) Multiple bar diagram : It is used when two or more related sets are too explained.

Here for each value of an attribute, we draw a pair of rectangles; each represents the value of respective sets.

Problem : Represent the following data in multiple bar diagram

Blood Group	Male	Female
A	15	15
B	25	15
AB	10	15
O	5	10



(iii) Proportional bar diagram : It is Alternative to multiple bar diagram.

There are two or more sets representing single attribute.

But the main difference is that in percentage bar diagram the height of bar is kept same i.e. 100% and each value is calculated in percentage.

Problem : Represent the following data in percentage bar diagram

Blood Group	Male	Female	Total
A	15	15	30
B	25	15	40
AB	10	15	25
O	5	10	15

Solution :

(i) For Blood group A-

Male % $= \frac{100}{30} \times 15 = 50\%$ Female % $= \frac{100}{30} \times 15 = 50\%$

(ii) For Blood group B-

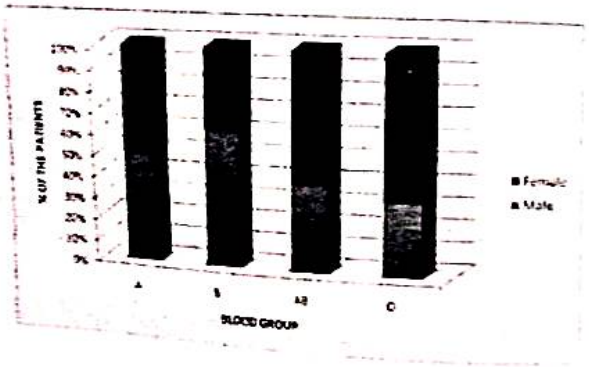
Male % $= \frac{100}{40} \times 25 = 62.5\%$ Female % $= \frac{100}{40} \times 15 = 37.5\%$

(iii) For Blood group AB-

Male % $= \frac{100}{25} \times 10 = 40\%$ Female % $= \frac{100}{25} \times 15 = 60\%$

(iv) For Blood group O-

Male % $= \frac{100}{15} \times 5 = 33.33\%$ Female % $= \frac{100}{15} \times 10 = 66.66\%$



(2) Pie diagrams : It presents discrete data of qualitative characters.

The frequencies of the groups are shown in a circle.

A degree of angle denotes the frequency and area of sector.

It gives comparative difference at a glance.

The Size of each angle calculated by the formula –

$$= \frac{Class\ frequence}{Total\ Observation} \times 360$$

Ex– (i) Distribution of cancer according to site is given below. Draw pie-chart

Lungs – 30%

Pancreas –10%

Stomach – 5%

Oral cavity – 50%

Penis – 5%

Solution : First we have to calculate the angles from given data

(i) Lungs – 30% i.e. $\frac{30}{100} \times 360 = 108$

(ii) Pancreas – 10% i.e. $\frac{10}{100} \times 360 = 36$

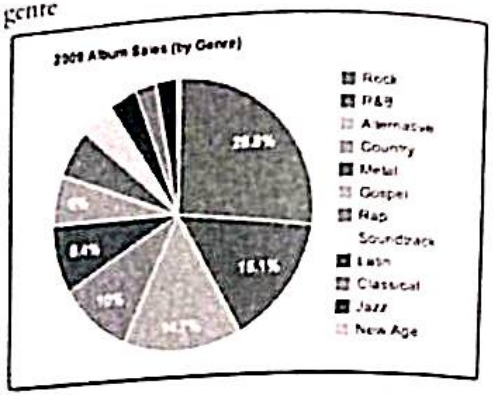
(iii) Stomach – 5% i.e. $\frac{5}{100} \times 360 = 18$

(iv) Oral cavity – 50% i.e. $\frac{50}{100} \times 360 = 180$

(v) Penis – 5% i.e. $\frac{5}{100} \times 360 = 18$

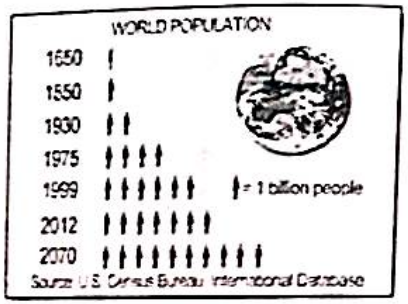
Type of Cancer	Value	Angle
Lungs	30%	108°
Pancreas	10%	36°
Stomach	5%	18°
Oral cavity	50%	180°
Penis	5%	18°

(ii) Ex – : pie chart depicting 463,124 total 2009 album sales broken into categories by genre



(3) Pictogram or Picture diagram : It is a popular method to impress the frequency of occurrence of events to common man. In this the frequencies or observations are shown in picture formant.

Ex – (i) Pictograph depicting population growth from 1650 into 2070 (projected)



(ii) A nurse is collecting blood type data from her patients. When a new patient is checked in, the nurse does a simple finger-prick test to see whether the patient's blood is type A, B, AB, or O.She tracks her results by creating a two-column table with the patient's name and blood type.

Name	Blood Group
Jayesh	A
Rupali	O
Rahul	AB
Madan	O
Pradip	AB

Name	Blood Group
Saurav	B
Jagadish	O
Baban	O
Pandit	B
Lata	AB

Solution :

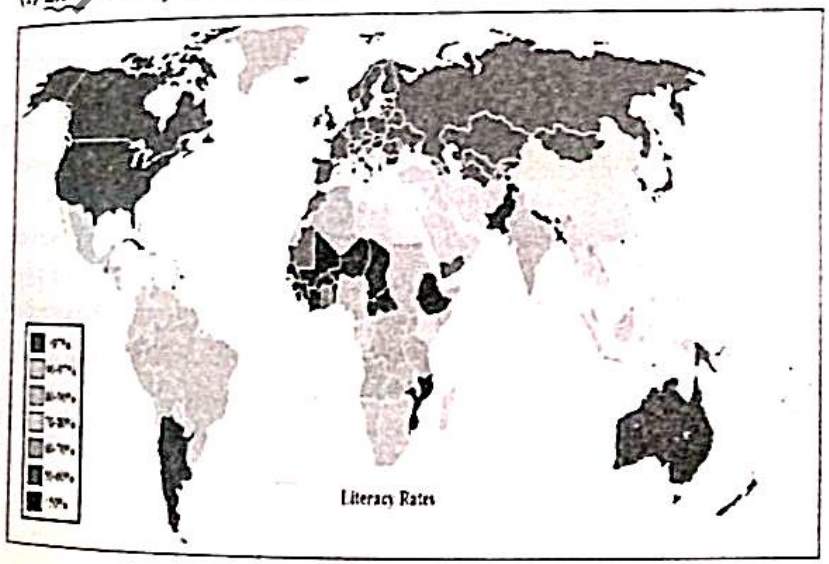
Blood Type	Number of People
Type A	💧
Type B	💧💧
Type AB	💧💧💧
Type O	💧💧💧💧

💧 = 1 person

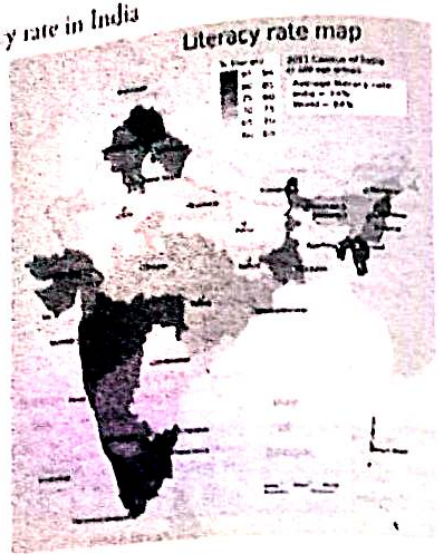
(4) Map or spot diagrams : It shows geographical distribution of frequency of characters.

It is especially useful while denoting the vital statistics of area.

(i) Ex – Literacy rates of world



(iii) *Ex* – literacy rate in India



(II) Graphs

(a) Histogram : It is a Graphical presentation of frequency distribution.

It is an area diagram.

The area of rectangle is varies with the frequency.

The variable characters of different groups are indicated on horizontal line (x - axis) which is called as "Abscissa" and the Frequency marked on vertical line (y axis) is called as "Ordinate". Such type of diagram is known as Histogram.

The height of rectangle alone will indicate the Frequency if a class interval is uniform.

The Area of rectangle alone will indicate the Frequency, if class intervals are different.

Histogram is a series of consecutive bars used for presenting Frequency distribution of grouped data. There is no width between two bars denoting that classes are continuous.

The width of bar is equal to 1cm & height is equal to frequency of the class given that all classes have equal class width.

Frequency = width of rectangle (bar) × height of rectangle.

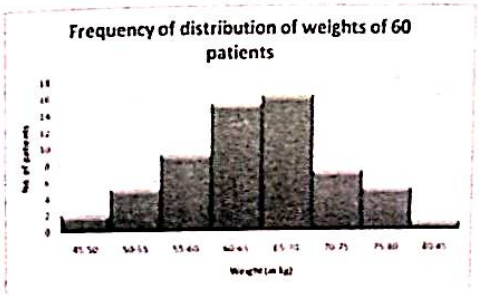
When the frequency distribution consists of unequal classes, we have to take into consideration 'frequency density' of classes and plot accordingly.

$$Frequence\ of\ density = \frac{frequency\ of\ class}{class\ width}$$

In short, if grouped frequency distribution is not continues, first it is to be converted into continues distribution and then histogram is drawn.

Ex – Frequency distribution of weights of 60 patients before they undergo basti treatment are given below. Draw a histogram.

Weight (in kg)	Number of patients (*f*)
45-50	2
50-55	5
55-60	9
60-65	15
65-70	16
70-75	7
75-80	5
80-85	1
Total N=	60



(b) Frequency Polygon : Polygon is a figure with many angles.

It is used when sets of data are to be illustrated on same diagram such as birth rates, death rates etc.

Ex – weight of patients before and after basti treatment can be explained using two frequency polygons.

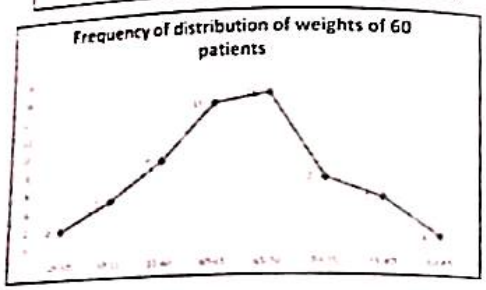
It is obtained either by joining the midpoints of class intervals at the height of frequencies by straight lines or by plotting frequencies of different classes directly against midpoint of classes.

Here X – Axis indicates the Categories of data and

Y – Axis indicates the Frequency of data in each category.

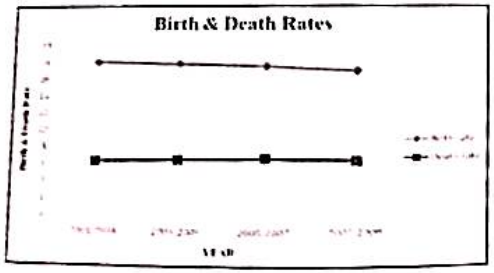
Ex – Frequency distribution of weights of 60 patients before they undergo basti treatment are given below. Draw a histogram.

Weight (in kg)	Number of patients (f)
45-50	2
50-55	5
55-60	9
60-65	15
65-70	16
70-75	7
75-80	5
80-85	1
Total N =	60



(ii) Ex – birth & death rates of state Maharashtra are given below. Draw frequency polygon.

Health Parameter	Year			
	2001-2003	2003-2005	2005-2007	2007-2009
Birth rate	18.1	17.9	17.6	17.1
Death rate	6.4	6.6	6.7	6.5

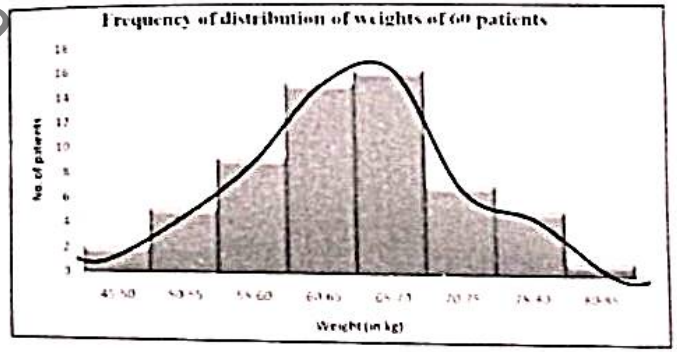


(c) **Frequency Curve :** When numbers of observations are very large and group interval reduced then the Frequency polygon tends to loss its angulation giving rise to smooth curve called as 'Frequency curve.'

Curve obtained in Normal distribution of individuals in large sample OR Normal distribution of means in population OR Normal distribution of difference in pairs of sample mean.

Ex – Frequency distribution of weights of 60 patients before they undergo basti treatment are given below. Draw a histogram.

Weight (in kg)	Number of patients (f)
45-50	2
50-55	5
55-60	9
60-65	15
65-70	16
70-75	7
75-80	5
80-85	1
Total N =	60



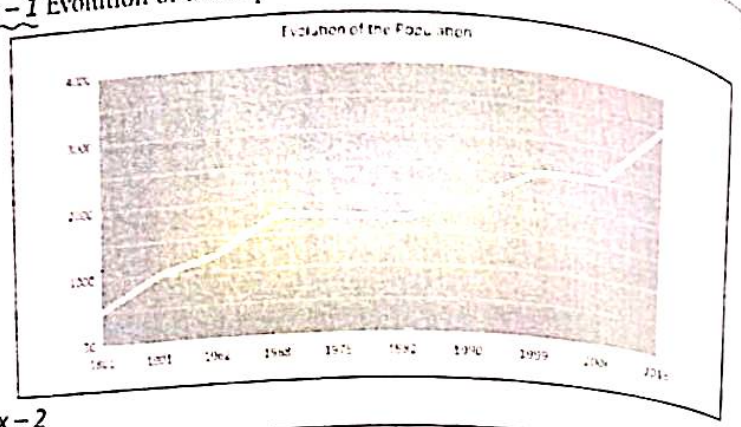
(d) Line Charts : This is a frequency polygon presenting variation by line.

It shows the trend of an event occurring over period of time rising, falling and showing fluctuation.

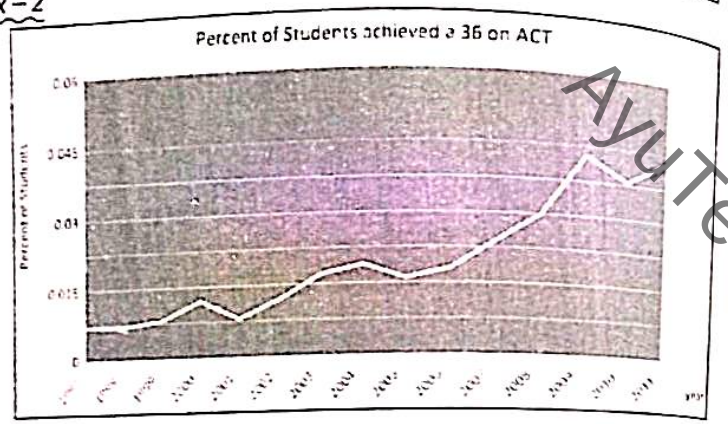
The class interval may be in months or year or 5years or10 years.

The shape of line chart may alter with change of scale on X or Y axis but trend indicated remains the same.

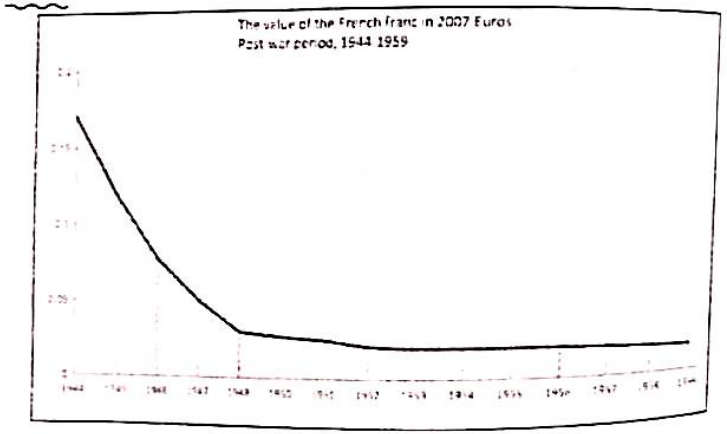
Ex – 1 Evolution of the Population.



Ex – 2



Ex – 3

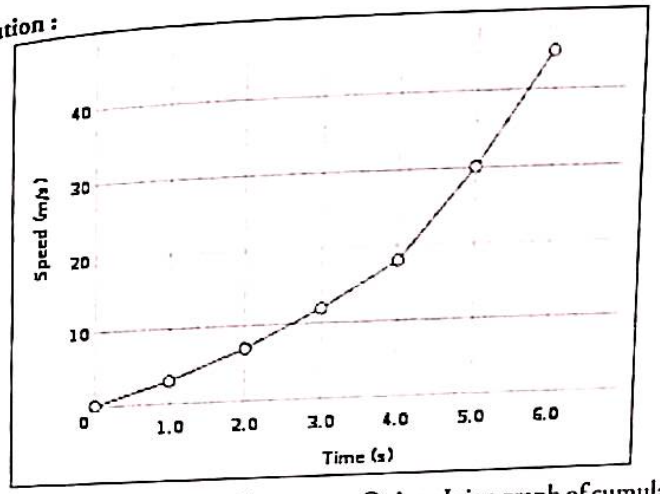


Ex – 4 If one were to collect data on the speed of a body at certain points in time, one could visualize the data by a data table such as the following:

Graph of Speed Vs Time

Elapsed Time (s)	Speed ($m\ s^{-1}$)
0	0
1	3
2	7
3	12
4	20
5	30
6	45.6

Solution :



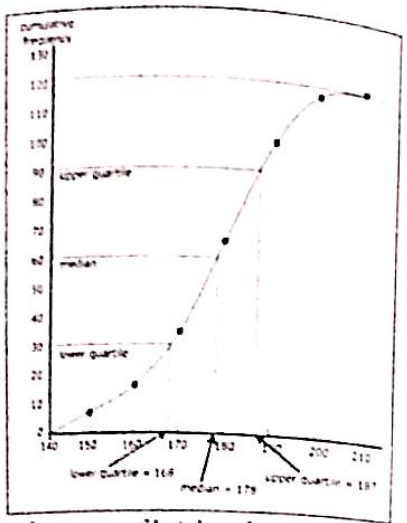
(e) Cumulative Frequency diagram or Ogive : It is a graph of cumulative relative frequency distribution.

An ordinary frequency distribution table in quantitative data has to be converted into relative cumulative frequency table.

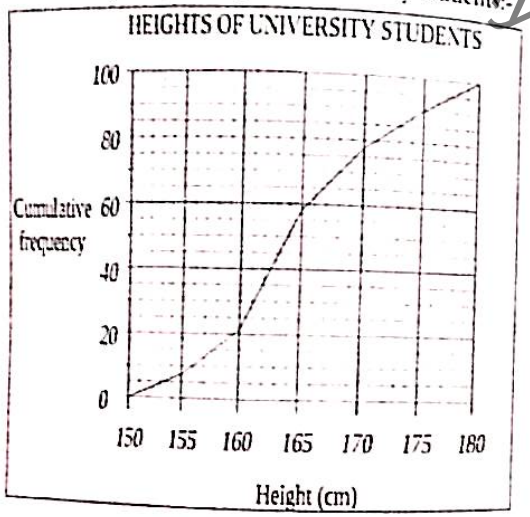
Then these frequencies are plotted corresponding to group limits of characteristics.

On joining the points by smooth free hand a curve diagram made is called as 'Ogive.'

Ex – 1



Ex – 2 Cumulative frequency of height of university students:-



(f) Scatter or Dot diagram : It is a graphical presentation made to show the nature of correlation between two variables X and Y in same person or group, hence it is also called as "co-relation diagram".

It is prepared after tabulation in which the frequencies of at least two variables have been cross classified.

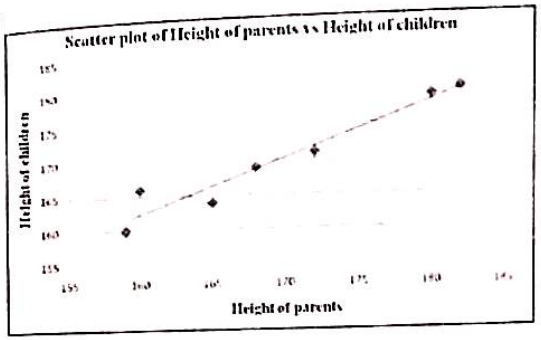
It is extremely useful in regression analysis.

Varying frequencies of characters give the number of such points or dots that shows a scatter line which is drawn to show the nature of correlation at a glance.

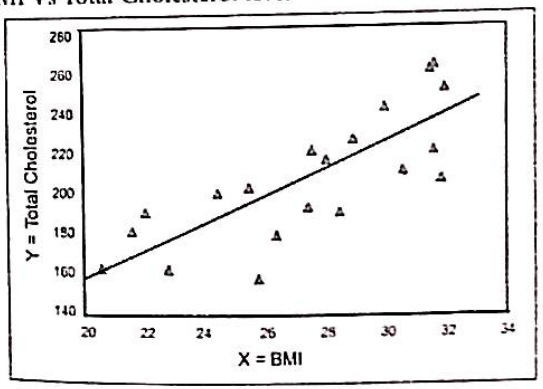
Ex – 1 Height of parents' vs Height of children given, Draw a scatter plot diagram.

Height of parents (X)	Height of children (Y)
165	164
160	166
180	179
182	180
172	171
168	169
159	160

Solution :



Ex – 2 -BMI Vs Total Cholesterol level-



QUESTIONS

1. Explain sources of data
2. Write in details presentation of data
3. Short notes on-
   Tabular presentation
   Drawing presentation.
4. Explain following term with example-
   (a) Histogram
   (b) Frequency polygon
   (c) Frequency curve
   (d) Line chart
   (e) Cumulative frequency
   (f) Scatter or dot
   (g) Bar diagram
   (h) Pie diagram
   (i) Picture diagram
   (j) Map/spot diagram

• • •

Chapter 4

MEASURES OF LOCATION

CHAPTER OVERVIEW

- Measures of Location : —
  - Average
  - Percentile
- Measures of Central Tendency
  - Mean
  - Mode
  - Median

Often it is not possible to list all the data or draw a histogram; it would be nice to have one number which best represents a data set. Often where the data lies is of interest, for which purpose a measure of location is useful.

What is Location : A fundamental task in many statistical analyses is to estimate a location parameter for the distribution; i.e., to find a typical or central value that best describes the data.

Definition : A single number used to characterize a batch, sample, or distribution or the position of an element of it.

In short location suggests the actual position of that point in the whole data. This position may be located either at center of line or nearby that center, which suggest the term average or percentile.

The Average measures a central value; therefore they located at the center or midpoint of distribution. But it may also be of interest to locate other point in the range. So the Percentile gives these values. They are values of variable such as height, weight, age etc.

AVERAGE : There are several different meanings of "average". The most generalis a "measure of central tendency", meaning any statistic that in some sense represents a typical value from a data set. The mean, median, and mode are often identified as "averages" in this sense.

The word "average" is also used (especially at elementary levels) to refer specifically to the mean, which is the kind of average you mentioned: add the numbers and divide by how many there are. This kind of average has

a specific meaning: it is the number you could use in place of each of the values, and still have the same sum.

For example, I'll illustrate the idea by making several piles of, say, beans. Suppose I make 5 piles, containing 4, 10, 9, 6, and 11 respectively. If I wanted to redistribute them into five piles each of which had the same number, I would gather them all together, count them (4+10+9+6+11 = 40), and then divide them evenly into 5 piles of 8 (40 / 5 = 8). Thus, the average is the number I get when I distribute a sum evenly; it smooths out the variations in the numbers.

Definition : (I) In statistics, an average is defined as the number that measures the central tendency of a given set of numbers.

A calculated "central" value of a set of numbers is called as Average. It is calculated by adding all the numbers, and then divided by how many numbers there are.

Ex – what is the average of 2, 7 and 9?

Solution :

i. Add the numbers: 2 + 7 + 9 = 18

ii. Divide by how many numbers (i.e. we added 3 numbers): 18 / 3 = 6

iii. So the average is 6.

(II) A number expressing the central or typical value in a set of data, in particular the mode, median, or (most commonly) the mean, which is calculated by dividing the sum of the values in the set by their number.

PERCENTILE : Percentiles report the relative standing of a particular value within a statistical data set. If that's what you're most interested in, the actual mean and standard deviation of the data set are not important, and neither is the actual data value. What's important is where you stand— not in relation to the mean, but in relation to everyone else. That's what a percentile gives you.

For example, in the case of exam scores, who cares what the mean is, as long as you scored better than most of the class? Who knows, it may have been an impossible exam and 40 points out of 100 was a great score. In this case, your score itself is meaningless, but your percentile tells you everything.

Suppose your exam score is better than 90% of the rest of the class. That means your exam score is at the 90th percentile (so k = 90), which hopefully gets you an A grade. Conversely, if your score is at the 10th percentile, then k = 10; that means only 10% of the other scores are below yours, and 90% of them are above yours; in this case an A grade is not in your future.

A nice property of percentiles is they have a universal interpretation: Being at the 95th percentile means the same thing no matter if you are looking at exam scores or weights of packages sent through the postal service; the 95th percentile always means 95% of the other values lie below yours, and 5% lie above it. This also allows you to fairly compare two data sets that have different means and standard deviations (like ACT scores in reading versus math). It evens the playing field and gives you a way to compare apples to oranges, so to speak.

A percentile is not a percent; a percentile is a value (or the average of two values) in the data set that marks a certain percentage of the way through the data. Suppose your score on the Final Exam was reported to be the 80th percentile. This doesn't mean you scored 80% of the questions correctly. It means that 80% of the students' scores were lower than yours and 20% of the students' scores were higher than yours.

Definitions : They are values in series of observations arranged in an increasing order of magnitude which divides distribution into 100 equal parts. *Or*

Any of the ninety-nine points that divide an ordered distribution into one hundred parts, each containing one per cent of the population. Or

Any one of the hundred groups so divided.

Ex – He is highly intelligent – he has an IQ in the top percentile.

In short, the average shows the value towards the center and percentile or variability shows the value in other than center.

Quartile : Quartiles divide the ordered data into 4 equal parts.

There are 3 quartiles – Q_1, Q_2, and Q_3.

Q_1 separates lower 25% observations from higher 75% observations.

$$Q1 = \left(\frac{n+1}{4}\right)^{th} observation$$

Q_2 separates lower 50% observations from higher 50% observations. It always equal to median.

$$Q2 = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

Q_3 separates lower 75% observations from higher 25% observations.

$$Q3 = \left(\frac{3n+1}{4}\right)^{th} \text{observation}$$

Quintiles : Quintiles are 4 in number and divide distribution into 5 equal parts. The 20th percentile of first quintile will have 20% observations falling to its Left and 80% to its right.

Deciles : They are 9 in number and divide the distribution into 10 equal parts.

The 1st decile of 10th percentile will divide the distribution into 10% to left and 90% to right side while 9th decile divide the distribution into 90% to left and 10% to right.

The Median of variable can also be called as –

Second Quartile or

Fifth decile or

50th percentile.

Application and Use of Percentile –

1. It is a good measure of location.
2. It divides the frequency distribution into two parts.
3. Preparation of standard percentile such as quartile Q_1 or median Q_2 etc. for particular age, sexes etc.
4. It is used for comparison of one percentile value of a variable of one sample with that of another sample which is drawn from the same population or from different population.
5. It is used to study the growth in children.
6. It is used as a measure of dispersion.

MEASURES OF CENTRAL TENDENCY

Introduction : A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as

summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

What are the measures of central tendency? A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution.

There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

Some aspects of the data can be described quantitatively to represent certain features of the data. An average is one of such representative measures. As average is a number of indicating the representative or central value of the data, it lies somewhere in between the two extremes. For this reason, average is called a measure of central tendency.

In this lesson, we will study some common measures of central tendency:

(i) Arithmetical average, also called mean

(ii) Median

(iii) Mode

1. Mean

"The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average."

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capitol letter, pronounced "sigma", which means "sum of...":

$$\bar{x} = \frac{\sum x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as μ:

$$\mu = \frac{\sum x}{n}$$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

Definition :

Mean is defined as sum of all the observations and dividing it by total number of observations.

It is denoted by $\bar{X}$

(1) For Simple series –

$$\bar{X} = \frac{\sum xi}{n}$$ Where x_i = observations

n = Total number of observations.

Ex – 1

Given data – 5, 5, 6, 7, 7, 7, 8, 9, 0, 8.

Solution -

Here $n = 10$

Mean $\bar{X} = \frac{\sum xi}{n} = \frac{5+5+6+7+7+7+8+9+0+8}{10} = \frac{62}{10}$

$$\therefore \bar{X} = 6.2$$

Ex – 2 Five people play golf and at one hole their scores are3, 4, 4, 5, 7for these scores, find – mean?

Solution -

Here $n = 5$

Mean $\bar{X} = \frac{\sum xi}{n} = \frac{3+4+4+5+7}{5} = \frac{23}{5}$

$$\therefore \bar{X} = 4.6$$

Ex – 2 Theretirement age distribution is given; find mean?

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

Solution -

Here n = 11

Mean $\bar{X} = \frac{\sum xi}{n} = \frac{54+54+54+55+56+57+57+58+58+60+60}{11} = \frac{623}{11}$

$$\therefore \bar{X} = 56.6$$

Ex – 2 The weights of four bags of wheat (in kg) are 103, 105, 102, and 104; Find themean weight.

Solution:

Mean weight $(\bar{X}) = \frac{103+105+102+104}{4}$ kg $= \frac{414}{4}$ kg

$$\therefore \bar{X} = 103.5 \text{ kg}$$

For Discrete series –

$$(\bar{X}) = \frac{\sum fixi}{N}$$

Where f_i – frequency.

N – Total frequency.

Ex – 1 Data given –

Number of accidents (x_i)	Number of days (f_i)
00	14
01	21
02	25
03	43
04	51
05	40
06	39
07	12

Solution :

x_i	f_i	$f_i x_i$
00	14	00
01	21	21
02	25	50
03	43	129
04	51	204
05	40	200
06	39	234
07	12	84
	N=245	922

As $(\overline{X}) = \frac{\sum fixi}{N} = \frac{922}{245} = 3.763 \approx 4$

$\therefore$ Mean $\overline{X} = 4$ accidents/day

Ex – 2 A football team keeps records of the number of goals it scores per match during a season. The list is shown opposite. Find the mean number of goals per match.

No. of Goals	Frequency
0	8
1	10
2	12
3	3
4	5
5	2

Solution : The previous table can be used, with a third column added. The mean can now be calculated.

No. of Goals (x_i)	No. of Frequency (f_i)	Goals × Frequency ($f_i x_i$)
0	8	0 × 8 = 0
1	10	1 × 10 = 10
2	12	2 × 12 = 24
3	3	3 × 3 = 9
4	5	4 × 5 = 20
5	2	5 × 2 = 10
TOTALS	40	73

As $\overline{X} = \frac{\sum fiXi}{N} = \frac{73}{40} = 1.825$

$\therefore$ Mean $\overline{X} = 1.825$

Ex – 3 The following data represents the weekly wages (in rupees) of the employees:

Weekly wages (in rupees)	900	1000	1100	1200	1300	1400	1500
Number of employees	12	13	14	13	14	11	5

Find the mean weekly wages of the employees?

Solution :

Weekly wages (in rupees) (x_i)	Number of employees (f_i)	($f_i x_i$)
900	12	10800
1000	13	13000
1100	14	15400
1200	13	15600
1300	12	15600
1400	11	15400
1500	5	7500
	Σf_i = N = 80	$\Sigma f_i x_i$ = 93300

As $\overline{X} = \frac{\sum fiXi}{N} = \frac{93300}{80} = 1166.25$

$\therefore$ Mean weekly wages $\overline{X} = 1166.25$

(3) For Continuous series –

$$\bar{X} = \frac{\sum fiXi}{N}$$

X_i = midpoint of classes.
N = Total frequency.

Ex – 1

Age group	Number of Heart surgeries (f_i)
10-20	01
20-30	02
30-40	15
40-50	18
50-60	17
60-70	10
70-80	07

Solution :

Age group	(f_i)	X_i (midpoints)	fX_i
10-20	01	15	15
20-30	02	25	50
30-40	15	35	525
40-50	18	45	810
50-60	17	55	935
60-70	10	65	650
70-80	07	75	635
N =	70		3570

$$\bar{X} = \frac{\sum fiXi}{N} = \frac{3570}{70} = 51$$

$\therefore$ Mean $\bar{X}$ = 51

Ex – 2 Find the mean for the following frequency distribution

Class	Frequency
20-40	9
40-60	11
60-80	14
80-100	6
100-120	8

Class	Frequency
120-140	15
140-160	12
Total	7

Solution :

Class	Frequency (f_i)	Class marks (X_i)	f_iX_i
20-40	9	30	270
40-60	11	50	550
60-80	14	70	980
80-100	6	90	540
100-120	8	110	880
120-140	15	130	1950
140-160	12	150	1800
	Σf_i = N = 75		Σf_iX_i = 6970

$$\bar{X} = \frac{\sum fiXi}{N} = \frac{6970}{75} = 92.93$$

$\therefore$ Mean $\bar{X}$ = 92.93

Ex – 3

Consider the following grouped frequency distribution:

Daily wages (in rupees)	Number of workers
150-160	5
160-170	8
170-180	15
180-190	10
190-200	2

Solution :

Daily wages (in rupees)	Number of workers (f_i)	Class marks (X_i)	f_iX_i
150-160	5	155	775
160-170	8	165	1320
170-180	15	175	2625
180-190	10	185	850
190-200	2	195	390
	Σf_i = 40		Σf_iX_i = 6960

$$\bar{X} = \frac{\sum fiXi}{N} = \frac{6960}{40} = 174$$

$\therefore$ Mean $\bar{X}$ = 174

Merits

Easy to calculate

Easy to understand

Rigidly defined

It is Available for further mathematical treatment.

It is based on all observations.

It is Least affected by sampling fractions.

Demerits

Only used by quantitative data.

It cannot determine graphically.

It is affected by extreme observations.

It can't obtain even if single observation is missing.

(2) MODE – Mode is also one of the measures of central tendency. The observation that occurs most frequently in the data is called mode of the data.

In other words, the observation with maximum frequency is called mode of the data.

Ex – The readymade garments and shoe industries etc. make use of this measure of central tendency. Based on mode of the demand data, these industries decide which size of the product should be produced in large numbers to meet the market demand.

Definition – An Observations having maximum frequency i.e. (observations occurring maximum number of times) called as MODE.

It is rarely used in medical studies.

For Simple series – An observation which occurs maximum number of times said to be Mode of that data.

Ex – The numbers of goals scored by a football team in 12 matches are:1, 2, 2, 3, 1, 2, 2, 4, 5, 3, 3, 4; what is the modal score?

Solution:– Just by looking at the data, we find the frequency of 2 is 4 and is more than thefrequency of all other scores.

So, mode of the data is 2, or modal score is 2.

Ex–2 In a survey of 10 households, the number of children was found to be 4, 1, 5, 4, 3, 7, 2, 3, 4, and 1. Find mode

Solution:– Given data: - 4, 1, 5, 4, 3, 7, 2, 3, 4, 1

As we know the observation occurring at maximum number of times known as Mode, hence the Mode = 4.

Ex – 3 In a beauty contest, the scores awarded by eight judges were:5.9, 6.7, 6.8, 6.5, 6.7, 8.2, 6.1, 6.3, so Using the eight scores, determine mode

Solution : – So from above data, we find the frequency of 6.7 is more than thefrequency of all other scores.

So, mode of the data is 6.7.

(2) For Discrete series – The Value of observation which is occurring maximum number of times.

Ex–1

No. of accidents	No. of days
0	14
1	21
2	25
3	43
4	51
5	40
6	39
7	12

In this 4 is having maximum number of frequency, hence mode is 4.

Mode = 4

Ex – 2 Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60 Find mode

Solution:-

Age	Frequency
54	3
55	1

Age	Frequency
56	1
57	2
58	2
60	2

The most commonly occurring value is 54; therefore the mode of the distribution is 54 years.

(3) For Continuous series –

$$\text{Mode} = l + \frac{f1 - f0}{2f1 - f0 - f2} \times h$$

l = lower class limit of modal class
f_1 = freq. of modal class
f_0 = freq. of pre modal
f_2 = freq. of post modal
h = class width

Ex – 1 Alex did a survey of how many games each of 20 friends owned and got this data, find mode?

Number of games	Frequency
1 - 5	2
6 - 10	7
11 - 15	8
16 - 20	3

Solution : We can easily find the modal group (the group with the highest frequency), which is 11 - 15

So we can say "the modal group is 11 - 15"

We can estimate the Mode using the following formula:

$$\text{Estimated Mode} = l + \frac{f1 - f0}{2f1 - f0 - f2} \times h$$

l = lower class limit of modal class
f_1 = freq. of modal class
f_0 = freq. of pre modal
f_2 = freq. of post modal
h = class width

In this example:
l = 11, f_0 = 7, f_1 = 8, f_2 = 3, h = 5

$$\text{Estimated Mode} = 11 + \frac{8 - 7}{(8 - 7) + (8 - 3)} \times 5$$

$$= 11 + (1/6) \times 5$$

$$\text{MODE} = 11.833$$

Ex – 2

Age	No. of Surgeries
10-20	1
20-30	2
30-40	15
40-50	18
50-60	17
60-70	10
70-80	7
Total	70

Solution : Firstly we have to choose the modal class which have maximum number of frequency ∴ Modal class is = 40 – 50 age group

$$\text{Mode} = l + \frac{f1 - f0}{2f1 - f0 - f2} \times h$$

Where - l = lower class limit of modal class = 40
f_1 = freq. of modal class = 18
f_0 = freq. of pre modal = 15
f_2 = freq. of post modal = 17
h = class width = 10

$$= 40 + \frac{18 - 15}{2 \times 18 - 15 - 17} \times 10$$

$$= 40 + \frac{18 - 15}{36 - 15 - 17} \times 10$$

$$= 40 + \frac{3}{4} \times 10$$

$$= 40 + 0.75 \times 10$$

$$= 40 + 7.5$$

$$= 47.5$$

MODE = 47.5

Ex – 3 The ages of the 112 people who live on a tropical island are grouped as follows, find Mode?

Age	Number
0 - 9	20
10 - 19	21
20 - 29	23
30 - 39	16
40 - 49	11
50 - 59	10
60 - 69	7
70 - 79	3
80 - 89	1

Solution : A child in the first group 0 - 9 could be almost 10 years old. So the midpoint for this group is 5 not 4.5 and hence the midpoints are 5, 15, 25, 35, 45, 55, 65, 75 and 85.

So, in the calculations of Mode, we will use the class boundaries 0, 10, 20 etc.

The Modal group is the one with the highest frequency, which is 20 - 29

l = 20 (the lower class boundary of the modal class)

f_0 = 21

f_1 = 23

f_2 = 16

h = 10

$$\text{Estimated Mode} = 20 + \frac{23 - 21}{(23 - 21) + (23 - 16)} \times 10$$
$$= 20 + 2.22$$
$$= 22.2$$

Merits –

1. Easy to calculate and understand.
2. Doesn't affect by extreme observations.
3. Calculated graphically using histogram.

Demerits -

1. Ill – defined i.e. more than 1 mode is exists.

2. Not based on all observations.
3. Not suitable for further mathematical calculations.

Empirical formula – If more than one mode exists.

Mean – mode = 3 (Mean - median)

(3) MEDIAN - When all observations of variable arranged either increasing or decreasing order, the middle observation is called as median.

It divides observations into 2 equal parts i.e. one less than median and other greater than median.

For Simple series – Median of data is calculated as follows:

(i) Arrange the (numerical) data in an ascending (or descending) order

(ii) When the number of observations (n) is odd, the median is the value of $(\frac{n+1}{2})^{th}$ observation.

(iii) When the number of observations (n) is even, the median is the mean of the $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)$ observation.

In Brief :

For odd number of observation $= \left(\frac{n}{2}\right)^{th}$ observation

For even number of observation $= \frac{\left(\frac{n}{2}\right)^{th} \text{ observation} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$

Ex – 1 – **(For Odd Series)** The weights (in kg) of 15 dogs are as follows - 9, 26, 10, 22, 36, 13, 20, 20, 10, 21, 25, 16, 12, 14, and 19; Find the median weight.

Solution : Let us arrange the data in the ascending (or descending) order:

9, 10, 10, 12, 13, 14, 16, (19), 20, 20, 21, 22, 25, 36

Median

Here, number of observations = n = 15

So, the median will be $\left(\frac{n+1}{2}\right)^{th}$ observation. i.e.

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$
$$= \left(\frac{15+1}{2}\right)^{th} \text{observation}$$
$$= 8^{th} \text{ observation}$$

Median = 19kg.

Answer: - The median weight 19 kg conveys the information that 50% dogs have weightsless than 19 kg and another 50% have weights more than 19 kg.

Ex – 2 **(For even Series)** The points scored by a basketball team in a series of matches are asfollows:16, 1, 6, 26, 14, 4, 13, 8, 9, 23, 47, 9, 7, 8, 17 and 28; find the median of the data.

Solution : First we have to rearrange the data in ascending order i.e.

1, 4, 6, 7, 8, 8, 9, 9, 13, 14, 16, 17, 23, 26, 28, 47

Here number of observations = n =16

So, the median will be the mean of $\left(\frac{n}{2}\right)^{th}$ observation and $\left(\frac{n}{2}+1\right)^{th}$ observation

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} - \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

$$\text{Median} = \frac{\left(\frac{16}{2}\right)^{th} \text{observation} - \left(\frac{16}{2}+1\right)^{th} \text{observation}}{2}$$

$$= \frac{8^{th} \text{observation} - 9^{th} \text{observation}}{2}$$

$$= \frac{9+13}{2}$$

$\therefore$ Median = 11

Answer: - Here the median 11 conveys the information that the values of 50% of theobservations are less than 11 and the values of 50% of the observations are more than 11.

Ex – 3 Find the median of the following data, which gives the marks out of 15,obtained by 35 students in a mathematics test.

Marks obtained	3	5	6	11	15	14	13	7	12	10
Number of Students	4	6	5	7	1	3	2	3	3	1

Solution : First arrange marks in ascending order and prepare a frequency table as follows:

Marks obtained	3	5	6	7	10	11	12	13	14	15
Number of Students (frequency)	4	6	5	3	1	7	3	2	3	1

Here n = 35, which is odd. So, the median will be $\left(\frac{n+1}{2}\right)^{th}$ observation

i.e. $\left(\frac{n+1}{2}\right)^{th}$ observation $\left(\frac{35+1}{2}\right)^{th}$ observation = $\left(\frac{36}{2}\right)^{th}$ observation = 18^{th} observation.

To find value of 18^{th} observation, we prepare cumulative frequency table as follows:

Marks obtained	Number of students	Cumulative frequency
3	4	4
5	6	10
6	5	15
7	3	18
10	1	19
11	7	26
12	3	29
13	2	31
14	3	34
15	1	35

So from the above table, we see that 18^{th}observation is 7.

So, Median = 7

(2) For Discrete series –

Median = Minimum value having cumulative frequency *(c.f.)* Greater than $\left(\frac{n}{2}\right)$

Ex–1

No. of accident	Days	*c.f.*
0	14	14
1	21	35
2	25	60
3	43	103
4	51	154
5	40	194
6	39	233
7	12	245

Here $\frac{N}{2} = \frac{245}{2} = 122.5$

As we know, median class is that which having $c.f. > \frac{N}{2}$

$\therefore C.f.$ just more than 122.5 is 154. So median is 4

$\therefore$ MEDIAN = 4

(3) For Continuous series

Median is a value of Lowest class having *c.f.* more than $\frac{N}{2}$

$$\text{Median} = = l + \frac{\frac{N}{2} - c.f.}{f} \times h$$

Where – *l*- lower class limit of median class

N – Total frequency

c.f. - cumulative frequency of pre-median class

f – Frequency of median class

h –Class width of median class

Ex – 1

Age	No. of H. surgeries (*f*)
10-20	1
20-30	2
30-40	15
40-50	18
50-60	17
60-70	10
70-80	7
Total	70

Solution -

Age	(*f*)	Less than *c.f.* (*L.c.f.*)
10-20	1	01
20-30	2	03
30-40	15	18
40-50	18	36
50-60	17	53
60-70	10	63
70-80	7	70
Total	70	

Here $\frac{N}{2} = \frac{70}{2} = 35$

As we know, lowest class having *c.f.* more than Median class $\frac{N}{2} = 35$

So the class is 40 – 50.

$$\text{Median} = l + \frac{\frac{N}{2} - c.f.}{f} \times h$$

$$= 40 + \frac{35 - 18}{18} \times 10$$

$$= 40 + \frac{170}{18}$$

MEDIAN = 49.44

Ex – 2 The table below gives data on the heights, in cm, of 50 children. Find the median?

Class Interval	Frequency
140 - 150	6
150 - 160	16
160 - 170	21
170 - 180	9

Solution :

Class Interval	Frequency	Less than c.f. (l.c.f)
140 - 150	6	6
150 - 160	16	22
160 - 170	21	41
170 - 180	8	50

Here $\frac{N}{2} = \frac{50}{2} = 25$

As we know, lowest class having *c.f.* more than Median class $\frac{N}{2} = 25$

So the class is 160 – 170.

As we know the formula, $\text{Median} = l + \frac{\frac{N}{2} - c.f.}{f} \times h$

Here *l* = lower class limit of median class = 160,

$\frac{N}{2} = 25$,

c.f. =cumulative frequency of pre-median class = 22

f= Frequency of median class = 21

h = Class width of median class = 10

$$\therefore Median = l + \frac{\frac{N}{2} - c.f.}{f} \times 10$$

$$= 160 + \frac{25 - 22}{21} \times 10$$

$$= 160 + \frac{3}{21} \times 10$$

$$= 160 + 1.42$$

$$= 161.42$$

∴ Median = 161 (to the nearest cm)

Merits –

It is rigidly defined.

Easy to calculate and understand.

It doesn't affect by extreme observations.

It is calculated graphically as Intersection of Ogive curves.

Demerits –

Not based on all observations

Not suitable for further procedure.

SUMMARY : Mean, median and mode are the measures of central tendency.

(A) Mean : Mean is defined as sum of all the observations and dividing it by total number of observations and denoted by

(1) For Simple series –

$$\overline{X} = \frac{\sum xi}{n}$$

Where x_i = observations

n = Total number of observations.

(2) For Discrete series –

$$\overline{X} = \frac{\sum fixi}{N}$$

Where *fi* – frequency.

N – Total frequency.

(3) For Continuous series –

$$\overline{X} = \frac{\sum fixi}{N}$$

Xi = midpoint of classes.

N = Total frequency.

(B) MODE : It is the observations occurring maximum number of times.

(1) For Simple series – it is the valueof observation occurring maximum number of times.

(2)For Discrete series –The Value of observation which is occurring maximum number of times.

(3) For Continuous series –

$$Mode = l + \frac{f1 - f0}{2f1 - f0 - f2} \times h$$

Where *l* = lower class limit of modal class

f_1 = freq. of modal class

f_0 = freq. of pre modal

f_2 = freq. of post modal

h = class width

(4) Empirical formula – If more than one mode exists.

Mean – mode = 3 (Mean - median)

(C) MEDIAN : When all observations of variable arranged either increasing or decreasing order, the middle observation is called as Median.

(1) For Simple series –

For odd number of observation= $\left(\frac{n+1}{2}\right)^{th}$ observation

For even number of observation= $\frac{\left(\frac{n}{2}\right)^{th} observation + \left(\frac{n}{2}+1\right)^{th} observation}{2}$

(2) For Discrete series –

Median = Minimum value having cumulative frequency *(c.f.)* Greater than $\frac{N}{2}$

(3) For Continuous series –

Median is a value of Lowest class having *c.f.* more than $\frac{N}{2}$.

$$Median = l + \frac{\frac{N}{2} - c.f.}{f} \times h$$

Where – *l*- lower class limit of median class

N – Total frequency

c.f. - cumulative frequency of pre-median class

f – Frequency of median class

h – Class width of median class

Measures of location :

1. What are the measure of location explain in details
2. Explain in brief-
   (a) Average
   (b) Percentile
3. Write in details measures of central tendency with examples
4. Short note with example on-
   (a) Arithmetic Mean
   (b) Mode
   (c) median

• • •

Chapter 5
VARIABILITY AND ITS MEASURES

CHAPTER OVERVIEW

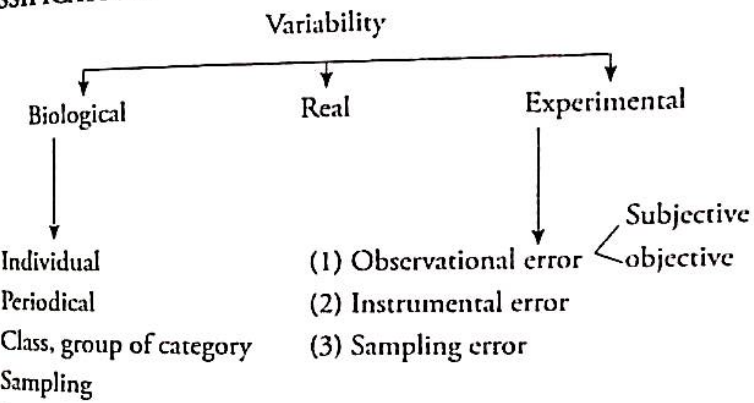
- Introduction
- Classification
- Percentiles
- Measures of Variability -
  (1) Range
  (2) Quartile Deviation
  (3) Mean Deviation
  (4) Variance
  (5) Standard Deviation
  (6) Coefficient Of Variation
- Standard Error

Variability

Variability is essentially a normal character and an occurrence of variability is a biological phenomenon.

Ex– no two measurements in man are absolutely equal, not even the means or proportions of two series in health or disease are equal though we compare like within like.

CLASSIFICATION



(1) BIOLOGICAL VARIABILITY – The variability occurring due to biological differences is known as biological variability.

It may be normal or natural difference in individual or group that occurs within certain accepted biological limits.

Ex-individuals in similar environments differ when compared as regards sex, class and other attribute. But the difference noted is may be small and it may occurred by chance.

Biological variability can be classified as –

(a) Individual variability

(b) Periodical variability

(c) Class, group or category variability

(d) Sampling variability

(a) Individual Variability – This variability is due to the fact that no two organisms, persons, patients are exactly equal. It varies from person to person.

Ex – One Students height is 160cm and of same age another Students height is 170cm

(b) Periodical Variability – The same individuals shows variations in temperature, BP, Pulse rate, BSL, Urea etc. at different times of day, in illness, during rest etc.

(c)) Class, group or category variability – Height, weight, BP etc. vary from class to class depending on age, sex, caste, social status etc.

(d) Sampling variability – To find mean or standard deviation of certain characteristics population sample is taken. If sample is small, then value of any sample will differ from those of population and there will be variability from one sample to another sample. This is a biological sampling error.

(2) REAL VARIABILITY – When the difference between two readings, observations or values of class or samples is more than the defined limits in universe, it is said to be real.

The cause may not be natural or inherent in man but it lies in external factor.

(3) EXPERIMENTAL VARIABILITY – Errors or difference or variation may be due to material, methods, procedures employed in study OR Defects in the techniques involved in experiment, known as Experimental Variability.

It is of three types –

(a) Observer error

(b) Instrumental error

(c) Sampling error.

(a) Observer Error –

(1) ***Subjective*** – Interviewer may alter some information thereby adding a number of errors while noting human peculiarities unless trained properly.

(2) ***Objective*** – errors may be added by an untrained observer while recording the measurements such as BP, pulse rate is known as Objective error.

(b) Instrumental Error – Defects in tools or apparatus or machine may cause undesirable variability OR error in observations leading to wrong conclusions and waste of money, time and labour is known as Instrumental error.

(c) Sampling Error – A sample drawn should not be biased or too small to draw conclusions. It should be representative and of sufficiently large sized to stand statistical tests.

Ex – Hospital based studies are mostly biased because the sample of patients under study is drawn from poor, influential or nearby strata of society. And many come to hospital at late stage.

MEASURES OF VARIABILITY

It is also called as "measures of desperation or variation or scatter."

Averages are best tool for summarizing the data but they have their own limitations. Hence variability is used.

Variability studied for –

(1) Finding out reliability of average.

(2) Determining cause and nature of variation and to control it.

(3) Compare two or more alternatives. Ex. – drug A and B.

Measures

Individual observations	Samples
(1) Range	(1) SE. of mean

(2) Quartile Deviation (Q.D.)

(3) Mean Deviation (M.D.)

(3) Standard Deviation (S.D.)

(5) Coefficient of Variation (C.V.)

(2) SE of different between two means

(3) SE of proportion

(4) SE of correlation coefficient

(5) S.D. of regression coefficient

(1) RANGE – R

Range is the simplest measure of dispersion.

It defines the limits of biological characteristics.

Definition : "The different between highest & lowest figures in given sample is known as Range."

It is not a satisfactory measure as it is based only on two extreme values, by ignoring all other observations within the extremes.

As we know the median & mode are not good measures of central tendency if compared with Mean, in the same way Range is also not a good measure of dispersion as compared to Standard deviation.

Range = Larger Value (L) – Smaller Value (S)

Merits –
(1)Rigidly defined
(2) It is Easy to calculate.

Demerits –
(1) Not based upon all observations.
(2) It can't be used with open ended class.
(3) Further mathematical calculations are not possible.
(4) It is affected by extreme observations.

(2) QUARTILE DEVIATION – The range of variable such as height between the first quartile (Q_1) and third quartile(Q_3) is called as "Inter-quartile range."

Median is the second quartile(Q_2) and half of this range is called as "Semi-quartile range."

Quartile Deviation is nothing but an average deviation between quartiles.

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2}$$

$$Q = \frac{Q_3 - Q_1}{2}$$

$$\therefore Q = \frac{Q_3 - Q_1}{2}$$

Merit –
(1) Easy to calculate
(2) Rigidly defined
(3) It doesn't affected by extreme observations.

Demerit –
(1) Not based on all observations.
(2) Further mathematical calculations are not possible.

(3) MEAN DEVIATION

Definition – "It is an average of absolute deviation of all observations from mean."

As we know the Quartile Deviation doesn't base on all observations, so to overcome this problem Mean Deviation is defined.

Steps – (1) Summing up of the differences from the mean.

$$|X_i - \bar{X}|$$

(2) Add the differences from the mean of each

$$\Sigma |X_i - \bar{X}|$$

(3) Divide the sum by number of observations

$$\text{M.D.} = \frac{\sum_{i=1} |X_i - \bar{X}}{n}$$

$$\therefore \text{Mean deviation about mean} = \frac{\sum X_i - \bar{X}}{n}$$

Merits –
(1) Rigidly defined
(2) Easy to calculate
(3) It is based on all observations.

Demerits – The further mathematical calculations are not possible

Question: A student took 5 exams in a class and had scores of 93, 65, 85, 90, and 78. Find the mean deviation for her test scores.

Solution:

Given:

n =5

First we have to calculate the Mean,

$\therefore$ Mean $\bar{X} = \frac{\sum X_i}{n}$ Where x_i = observations

n = Total number of observations.

$\therefore$ Mean $\overline{X} = \frac{93 + 65 + 85 + 90 + 78}{5} = \frac{411}{5} = 82.2$

$\therefore$ Mean $\overline{X}$ = 82.2

Now let's subtract the mean from each score, take the absolute value of each difference, total the absolute values and then divide by the number of values.

x_i	$x_i - \overline{X}$	$\lvert x_i - \overline{X} \rvert$
93	10.8	10.8
65	-17.2	17.2
85	2.8	2.8
90	7.8	7.8
78	-4.2	4.2
$\overline{X}$=82.2		$\Sigma X - \overline{X}$=42.8

So, Mean deviation $= \frac{\sum X_i - \overline{X}}{n}$

$= \frac{42.8}{5}$

$= 8.56$

$\therefore$ Mean deviation M.D.= 8.56

By this we can say that on the average, this student's test scores are deviated by 8.56 points from the mean.

(4) VARIANCE – S^2 As in Mean Deviation we ignored sign of deviation hence we can't use it for further mathematical calculations.

Definition – Variance is defined as "average of squared deviation of all observations from mean."

And is denoted by S^2

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

Demerit – As it is average squared deviation, its unit is not same as that of observations.

(5) STANDARD DEVIATION – Variance is the average squared deviation. Its unit is not same as that of observations. So to tackle this problem we took square root of it.

Definition – It is defined as - "the positive square root of arithmetic mean of the square."

Standard deviation is square root of variance and it is denoted by "σ" (sigma).

It measures scatter and estimate of scatter in population from which sample is drawn.

Steps :

(1) Calculate the mean – $(\overline{X})$

(2) Find difference of observation from the mean $(x_i - \overline{X})$

(3) Square different of observations From mean $(x_i - \overline{X})^2$

(4) Add the squared values - $\sum (x_i - \overline{X})^2$

(5) Divide this sum by number of observations to get variance.

$$S^2 = \frac{\sum (x_i - \overline{X})^2}{n}$$

To get Standard Deviation - find the square root of S^2

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n}}$$

Sometimes, only a sample of whole population is given. In this case instead of calculating standard deviationof population, we calculate standard deviationof sample. The formula for sample standard deviation is given below:

$$\textbf{SD of sample} = \sqrt{\frac{\sum (x_i - \overline{X})^2}{n-1}}$$

Uses –(1) It summarizes deviation.

(2) It indicates variation of different of observations from mean is by chance or real.

(3) It is used to find Standard Error.

(4) It helps to derive a suitable size of sample.

(5) It is rigidly defined.

(6) Further mathematical treatment is possible.

(8) It is not much affected by sampling fractions.

Question : During a survey, 6 students were asked that how many hours per day they study on an average? Their answers were as follows: 2, 6, 5, 3, 4, 1. Evaluate the standard deviation.

Solution : First we have to calculate the Mean,

Formula for mean is given by Mean $= \sqrt{\frac{\sum (x_i - \overline{X})^2}{n-1}}$

Where x_i = observations

n = Total number of observations.

$\therefore$ Mean $\bar{X} = \frac{2+6+5+3+5+1}{6} = \frac{21}{6} = 3.5$

Construct the following table for standard deviation:

x_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
2	-1.5	2.25
6	2.5	6.25
5	1.5	2.25
3	-0.5	0.25
4	0.5	0.25
1	-2.5	6.25
$\bar{X} = 3.5$		$(x_i - \bar{X})^2 = 17.5$

Formula for standard deviation is given by

$$S = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n}} = \sqrt{\frac{17.5}{6}} = \sqrt{2.92} = 1.71$$

$\therefore S = 1.71$

(6) COEFFICIENT OF VARIATION – It is used to compare relative variability.

It is calculated from Standard Deviation and mean of characteristic.

It is expressed as percentage (%) of mean.

$$CV = \frac{S.D}{Mean} \times 100$$

$$CV = \frac{s}{\bar{x}} \times 100$$

The ratio of Standard Deviation and mean is found in percentage.

Coefficient of Variation doesn't have unit so it is called as "Relative measure of dispersion."

More the CV – more is the variability.

It is used to compare variability of one character in two different groups having different magnitudes of values OR two characters in same group by expressing in percentage.

It has all the merits of Ideal measure of dispersion.

Ex – 1 The cholesterol level of 10 people is 260, 150, 160, 200, 210, 240, 220, 225, 210, and 240. Then Calculate -

i. Range

ii. Quartile Deviation

iii. Standard Deviation

iv. Coefficient of Variation.

Solution : Data is arranged in ascending order as- 150, 160, 200, 210, 210, 220, 225, 240, 240, 260.

$$\text{Mean } (\bar{X}) = \frac{150+160+200+210+220+225+240+240+210+260}{10} = 211.5$$

(1) Range

Range = Larger value – Smaller value

= 260 – 150

Range = 110

(2) Quartile Deviation -

$$Q.D. = \frac{Q3 - Q1}{2}$$

Q_1 = observation

$= (2.75)^{th}$ observation

$= 2^{nd}$ observation + 0.75 (3rd observation –– 2^{nd} observation)

= 160 + 0.75 (200 – 160)

$Q_1 = 190$ ---------- (i)

$Q_3 = \left(\frac{3n+1}{4}\right)^{th}$ observations

$= 7.75^{th}$ observation

$= 7^{th}$ observation + 0.75 (8th observation –– 7^{th} observation)

= 225 + 0.75 (240 – 225)

$Q_3 = 236.25$ ---------- (ii)

From (i) and (ii)

$$Q.D. = \frac{Q3 - Q1}{2} = \frac{236.25 - 190}{2} = \frac{46.25}{2}$$

Q.D. = 23.175

(3) Standard Deviation –

$$S = \sqrt{\frac{\Sigma(Xi - \bar{X})^2}{n}}$$

To calculate S.D. first we have to calculate following steps as –

X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
150	– 61.5	3782.25
160	– 51.5	2652.25
200	– 11.5	132.25
210	– 1.5	2.25
210	– 1.5	2.25
220	8.5	72.25
225	13.5	182.25
240	28.5	812.25
240	28.5	812.25
260	48.5	2352.25
	Total =	10802.5

As ,

$$S = \frac{\sqrt{\Sigma(X_i - \bar{X})^2}}{n}$$

$$= \frac{\sqrt{10802.5}}{10}$$

$$= \sqrt{1080.25}$$

$$S = 32.87$$

(4) Coefficient of Variation

$$C.V. = \frac{S.D.}{\bar{X}} \times 100$$

$$= \frac{32.87}{211.5} \times 100$$

$$C.V. = 15.5\ \%$$

Ex – 2 Two treatments A & B are tested on 10 patients each. The data of recovery time of treatment is given below. Find which treatment is more consistent.

Treatment	A	B
$\bar{X}$	8	10
SD	2	3

Solution :

$$CV \text{ for treatment A} = \frac{S.D.}{\bar{X}} \times 100$$

$$= \frac{2}{8} \times 100 \quad \text{.....(i)}$$

$$= 25$$

$$CV \text{ for treatment B} = \frac{S.D.}{\bar{X}} \times 100$$

$$= \frac{3}{10} \times 100 \quad \text{.....(ii)}$$

$$= 30$$

From (i) & (ii) –

CV for B is greater than CV for A i.e. 30>25

∴ Treatment A is more consistent than treatment B.

STANDARD ERROR

The standard deviation of sampling distribution of statistic is known as its 'Sampling Error' or "Standard Error" (SE) and it is measure of chance variation.

S.E. is a measure, which enables us to judge whether the mean of given sample is within set of confidence limits or not.

S.E. plays very important role in large sample theory and from the basis of testing of Hypothesis.

S.E. enables us to determine the probable limits within which population parameters may be expected to lie.

S.E. is especially important in dealing with statistics, which are normally distributed for such statistics; S.E. provides information concerning Probability that statistic will deviate from its parameter by specified amount.

S.E. is basis for test of significance of larger samples i.e.

S.E. of single mean or difference of two means.

S.E. of single proportion or different of two means.

Ex – S.E. of mean ($\bar{X}$) = $\frac{s}{\sqrt{n}}$

σ = S.D.

n = Sample size

S.E. of proportion (P) = $\sqrt{\frac{pq}{n}}$

Standard Error of Mean – [$S\overline{X}$]

It is calculated by drawing only one large sample and then its Standard deviation is calculated.

SE of $\overline{X}$ or $S\overline{X} = \frac{S.D.}{\sqrt{n}}$

$$S\overline{X} = \frac{S}{\sqrt{n}}$$

$$SE \propto S \;\&\; SE \propto \frac{1}{\sqrt{n}}$$

S.E. is minimized by reducing S.D. and S.D. is reduced by taking large sample.

Ex – Following are results of ESR in first hour observations in 5 persons.

Calculate S.E. – 2, 4, 6, 8, 10.

Solution

$$S.E. = \frac{S}{\sqrt{n}}$$

As we know,

$$S = \sqrt{\frac{\Sigma(Xi-\overline{X})^2}{n}}$$

So firstly we have to calculate $\overline{X}$,

$$\overline{X} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

x_i	$\overline{X}$	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$
2	6	-4	16
4	6	-2	4
6	6	0	0
8	6	2	4
10	6	4	16
Total			40

$$\therefore S = \sqrt{\frac{\Sigma(Xi-\overline{X})^2}{n}} = \sqrt{\frac{40}{5}} = \sqrt{8}$$

$S = 2.82$

Hence

$$S.E. = \frac{S}{\sqrt{n}} = \frac{2.82}{\sqrt{5}} = \frac{2.82}{2.23} = 1.26$$

S.E. = 1.26

Question 1: Calculate the standard error of the given data *x*: 10, 12, 16, 20, 24

Solution:

Given observations are,x: 10,12,16,20,24

Standard error formula is given by, SE of mean = $\frac{s}{\sqrt{n}}$

Mean can be calculated as,

$$\overline{X} = \frac{10+12+16+20+24}{5} = 16.4$$

Standard deviation can be calculated as, here we have to calculate the SD of given sample so we take (n-1) instead of n.

$$S = \frac{\sqrt{\Sigma(X_i-\overline{X})^2}}{n-1} = \frac{\sqrt{(10-16.4)^2+(12-16.4)^2+(16-16.4)^2+(20-16.4)^2+(24-16.4)^2}}{5-1}$$

$$= \frac{\sqrt{40.96+19.36+0.16+12.96+57.76}}{4}$$

$$= \frac{\sqrt{131.2}}{4}$$

$$= \sqrt{32.8}$$

$$= 5.73$$

∴**Standard Error= 2.56**

SE of mean $= \frac{s}{\sqrt{n}}$

$$= \frac{5.73}{\sqrt{5}}$$

$$= \frac{5.73}{2.23}$$

∴**Standard Error-2.56**

Question 2: Determine the standard error of the following observations.*x*: 5, 10, 15, 20, 25

Solution:

Given observations are, *x*: 5,10,15,20,25

Standard error formula is given by, SE of mean = $\frac{s}{\sqrt{n}}$

Mean can be calculated as,

$$\overline{X} = \frac{5+10+15+20+25}{5} = \frac{75}{5} = 15$$

Standard deviation can be calculated as, here we have to calculate the SD of given sample so we take (n-1) instead of n.

$$S = \frac{\sqrt{\Sigma(X_i - \bar{X})^2}}{n-1} = \frac{\sqrt{(5-15)^2+(10-15)^2+(15-15)^2+(20-15)^2+(25-15)^2}}{5-1}$$

$$= \frac{\sqrt{100+25+0+25+100}}{4}$$

$$= \frac{\sqrt{250}}{4}$$

$$= \sqrt{62.5}$$

$$= 7.90$$

$$\text{SE of mean} = \frac{s}{\sqrt{n}}$$

$$= \frac{7.90}{\sqrt{5}}$$

$$= \frac{7.90}{2.23}$$

∴ Standard Error= 3.54

QUESTIONS

1. explain variability & its measures
2. write classification of variability
3. explain percentile with example
4. write short notes on—
   (a) range
   (b) quartile deviations
   (c) mean deviations
   (d) standard deviations
   (e) variance
   (f) standard error

• • •

Chapter 6

PROBABILITY

CHAPTER OVERVIEW

- Introduction
- Definition
- Properties
- Types
  1. Mathematical Probability
  2. Frequentist Probability
  3. Bayesian Probability
- Laws of Probability
  1. Addition Law.
  2. Multiplication Law.
  3. Binomial Law.
  4. Probability From Shape of Normal Distribution.
  5. Probability of Calculated Values from Table.
- Test of Significance

INTRODUCTION

In day to day life, we tend to make statements about uncertain future events.

For example

1. There is a huge possibility that this vomiting is due to food poisoning.
2. I can't tell youthat this new drug is better than existing drug.
3. The success rate of this surgery is 90%.
4. It is confirmed that a TB patients getting DOTS treatment will get cured for sure.

These statements expresses the degree of certainty (and intuitively of uncertainty) of events. The term Probability is nothing but these certainties which are expressed in numerical format.

Before leading to mathematical definition of probability, let's have a look at the basic mathematical prerequisite.

Random Experiment : An Experiment is called random experiment if its 'exact' outcome cannot be known in advance & we can repeat this experiment in identical condition.

Ex–

Tossing a coin

Throwing a dice

Calling a roll number & noting down sex.

Trial of Experiment : Performing a random experiment is known as 'Trial of Experiment.'

Sample space : A set of all possible outcomes are known as 'Sample space of Experiment' & it is denoted by symbol "Ω".

Event : The outcomes or combinations of outcomes are known as 'Events' & are denoted by capital letters A, B, C etc.

It is clear that, the events are subsets of sample space.

TYPES OF EVENTS :

(1) Mutually Exclusive Event:-

Two or more events are called mutually exclusive events if they have no sample points in common. i.e. $A \cap B = \Phi$

(2) Complimentary Event : Let A be an event, then its complementary event is the set of all elements (sample points) of sample space except elements of A. it is denoted by A' or $\bar{A}$

Complimentary events are always mutually exclusive i.e. $A \cup \bar{A} = \Phi$

Also complimentary events areexhaustive. i.e. $A \cup \bar{A} = \Omega$

(3) Impossible Event : An event which doesn't contain any sample point from sample space is known as 'Impossible event.' It is denoted by an empty set.

$$A = \{\}$$

(4) Sure Event : An event which contains all the sample points from the sample space is known as 'Sure Event'.

i.e. if A is sure event, then $A = \Omega$

(5) Independent Event : To know about the independent events first we have to know about the term Conditional probability.

Conditional probability:- Let A & B be two events, then probability of A given that B has happened. This is known as 'Conditional Probability of 'A' & it is given by

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

When conditional probability of A i.e. $P\left(\frac{A}{B}\right)$ is equal to P (A), then these two events A & B are independent.

If $P\left(\frac{A}{B}\right) = P(A)$, then A & B are independent events.

Definition : "The relative frequency or probable chances of occurrence with which an event is expected to occur on an average."

OR

"It is Relative frequency with which an event is expected to occur on an average."

Ex.–

1. The twin pregnancy will occur once in 80 pregnancies.
2. A Child with blood group Rh negative will born once in 10 births.

Classical or Mathematical Definition : Let Ω be a sample space associated with random experiment, then probability of an event 'A' denoted by P (A) is given by

$P(A) = \frac{n(A)}{n(\Omega)}$

n (A) = number of favorable event.

$n(\Omega)$ = Number of executive event.

Properties : Probability usually expressed by symbol 'P'

1. It ranges from 0 to 1 i.e. $0 \leq P \leq 1$
2. P [sure event] = 1
3. P [impossible event] = 0
4. Probability of an event happening in sample is P and that of not happening is denoted by symbol q

Then –

$$P + q = 1$$

Types of Probability : The types of probability according to perspective and their interpretation –

1. Mathematical (Theoretical or prior)
2. Frequents (experimental or objective)
3. Bayesian (subtractive)

(1) Mathematical Probability : All possible outcomes are assigned equal Probability of happening i.e. occurrence of all outcomes are considered as equally likely.

$$P(A) = \frac{n(A)}{n(\Omega)}$$

Ex. – Tossing coin and find Probability Of outcome of head.

(2) **Frequentist Probability :** Probability interpreted as fraction of favorable cases (i.e. no. of times the event in question occurred) in large number of trials in the random experiment.

Ex. – standard setup of hypothesis testing based on this.

(3) **Bayesian Probability :** Probability of events that not only consider the output of trial but also consider investigators belief about Probability even before trial starts (prior).

Thus as prior believe subjects to differ according to different investigators, Probability is called "Subjective".

This Probability well explained using "Bayes' formula" -

$$P\left[\frac{Hypothesis}{Data}\right] = \frac{P\left(\frac{data}{Hypothesis}\right) \times P[Hypothesis]}{P[data]}$$

Example No. 1 Let a coin is tossed two times & outcomes are noted. What is the probability that we observe two heads?

Solution : Sample space is given by,

Ω = {HH, HT, TH, TT} where H – Head, T – Tails

Let A be the event that we observe 2 heads,

$\therefore A = \{HH\}$

$\therefore P(A) = \frac{n(A)}{n(\Omega)}$

$\therefore P(A) = \frac{1}{4}$

i.e. the probability that we observe two headsis $\frac{1}{4}$

Example No. 2 Let a dice is thrown & number on the upper surface is noted.

(i) What is the probability that the number is greater than 3?

(ii) What is the probability that the number is 8?

Solution : Sample space is given by,

$\Omega = \{1, 2, 3, 4, 5, 6\}$

(i) Let A be the event that number is greater than 3

$\therefore A = \{4, 5, 6\}$

$\therefore P(A) = \frac{n(A)}{n(\Omega)} = \frac{3}{6}$

$\therefore P(A) = \frac{1}{2}$

(ii) Let B be the event that number is 8

$\therefore B = \{\}$

$\therefore P(B) = \frac{0}{6}$

$\therefore P(B) = 0$

Hence (i) The probability of number is greater than 3 is

(ii) The probability of number is 8 is 0 (zero)

Example No 3 – 5 patients are waiting in OPD consisting of 3 men & 2 women. What is the probability that the first patient to be checked is male?

Solution : Here, the exhaustive number of cases i.e. the number of ways to choose a patient from 5 patients is 5.

$\therefore n(\Omega) = 5$

Let A be an event that the first patient to be checked is male & here the Number of ways to choose a patient from 3 men is 3.

$\therefore n(A) = 3$

Hence the probability that the first patient to be checked is male P (A) –

$\therefore P(A) = \frac{n(A)}{n(\Omega)} = \frac{3}{5}$

$\therefore P(A) = \frac{3}{5}$

LAWS OF PROBABILITY

i. Addition law.
ii. Multiplication law.
iii. Binomial law.
iv. Probability from shape of Normal distribution
v. Probability of calculated values from table.

(1) **ADDITION LAW** - Let A and B be two events with Probability P (A) and P (B) respectively.

Then Probability of event 'A OR B' is denoted by $(A \cup B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Note that A OR B means either A or B or both happens.

If A and B mutually exclusive then

$$P(A \cup B) = P(A) + P(B)$$

Ex – A fair dice is tolled and the number on the surface is noted. What is the probability that observed number is –

(i) multiple of 2

(ii) greater than 3

(iii) Multiple of 2 or greater than 3?

Solutions : Sample space is given as $\Omega = \{1, 2, 3, 4, 5, 6\}$

(i) Let 'A' be the event that we observe number is multiple of 2 i.e.

$$\therefore A = \{2,4,6\}$$

$$\therefore P(A) = \frac{n(A)}{n(\Omega)}$$

$$\therefore P(A) = \frac{3}{6}$$

$$\therefore P(A) = \frac{1}{2} \quad \text{---------- (I)}$$

(ii) Let 'B' be the event that we observe number is greater than 3 i.e.

$$\therefore B = \{4,5,6\}$$

$$\therefore P(B) = \frac{n(B)}{n(\Omega)}$$

$$\therefore P(B) = \frac{3}{6}$$

$$\therefore P(B) = \frac{1}{2} \quad \text{---------- (II)}$$

(iii) The event that number is multiple of 2 or greater than 3 is given by $(A \cup B)$.

So before that we have to calculate the event having number is multiple of 2 & greater than 3.

Let $(A \cap B)$ be an event that the number is multiple of 2 & greater than 3.

$$\therefore (A \cap B) = \{4, 6\}$$

$$\therefore P(A \cap B) = \frac{n(A \cap B)}{n(\Omega)}$$

$$\therefore P(A \cap B) = \frac{2}{6}$$

$$\therefore P(A \cap B) = \frac{1}{3} \quad \text{---------- (III)}$$

So according to law of probability, from equation (I), (II) and (III)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{1}{2} + \frac{1}{2} - \frac{1}{3}$$

$$= 1 - \frac{1}{3}$$

$$\therefore P(A \cup B) = \frac{2}{3}$$

i.e. the probability of multiple of 2 OR greater than 3 is $\frac{2}{3}$.

(2) MULTIPLICATION LAW -

Let A and B two Independent events, then the Probability of event 'A and B' is denoted by $P(A \cap B)$ and is given by product of their individual Probability.

$$P(A \cap B) = P(A) \times P(B)$$

This law is applied to 2 or more events occurring together but they must not be associated i.e. they must be independent of each other.

The word 'AND' is used in between the events & its sequence is immaterial.

Ex – Probability that a patient have blood group 'O' is $\frac{1}{10}$ and the probability of a person have HIV positive is $\frac{1}{10}$. What is the probability that a person having blood group 'O' is HIV positive?

Solution : Let 'A' be the event that a person is HIV positive,

$$\therefore P(A) = \frac{1}{10}$$

Let 'B' be the event that a person has blood group 'O',

$$\therefore P(B) = \frac{1}{10}$$

So now let $A \cap B$ be the event that a person is HIV positive AND has a blood group 'O'.

Note that event A and event B are independent

$$\therefore P(A \cap B) = P(A) \times P(B)$$

$$= \frac{1}{10} \times \frac{1}{10}$$

$$\therefore P(A \cap B) = \frac{1}{100}$$

(3) BINOMIAL LAW -

In any trial, if there are two possibilities – either 'success' or 'failure' such distribution is called as "Binomial Distribution."

Ex– Cured or not cured
Male or female
Failed or passed

It is discrete type of distribution and outcome of this will be in whole numbers.

For an Example –

When two children are born one after other, the possible sequence will be–

(1) 1st Male &2nd Male $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ i.e. 25%

(2) 1st Male &2nd Female $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ i.e. 25%

(3) 1st Female &2nd Male $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ i.e. 25%

(4) 1st Female &2nd Female $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ i.e. 25%

Here, chance of getting 2 Males $= \frac{1}{4}$ and Chance of getting 2 Females $= \frac{1}{4}$

So, Chance of getting 1 Male & 1 Female will be

= 1st Female & 2nd Male OR 1st Male & 2nd Female

$= \frac{1}{4} + \frac{1}{4}$

$= \frac{1}{2}$ (50%)

Now, if female child is born 1st and 2nd child is desired, then what is the probability that 2nd child will be male?

So, here we have to think about only 2nd one, not being a female child i.e. it means 2nd one is male child –

As P + q = 1

q = 1 – P

$= 1 - \frac{1}{4}$

$\therefore q = \frac{3}{4}$ (75%)

This mathematical calculation can be expressed by Binomial Expression as –

$$(P + q)^n$$

Where n = number of events or sample size

P = Probability of success

q = Probability of failure. (P + q) = 1

Ex– In case of 2 pregnancies or 2 events (n = 2)

$$(P + q)^n = (P + q)^2$$
$$= P^2 + 2Pq + q^2$$

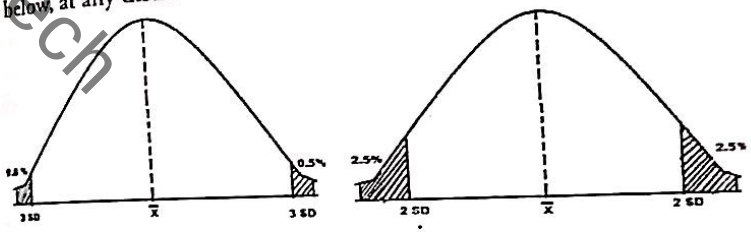
If P = male child & q = female child then

P^2 = both Childs were Male

Pq = one Male/ one female

p^2 = both Childs were Female

(4) Probability (chances) from Shape of Normal Distribution or Normal Curve – As we know the properties of normal distribution curve, the probability of any observation or number of observations lying above or below, at any distance from the mean can be estimated.



Ex–Heights of 200 individuals are taken. Then as per normal distribution –50% of individual will lie above the mean and 50% will lie below the mean.

Range of mean ± 1 SD = 68%

± 2 SD = 95% (rest 5%)

∴Probability of having height above the mean +2 SD is 2.5% & below the mean –2 SD is 2.5%

Similarly, the total area under the normal curve is taken as unity (one).

Proportional area under any part of curve will indicate relative frequency or probability of observations between any two points.

Ex – Probability of observation falling beyond ± 2 SD is very small i.e. 2.5 + 2.5 = 5% only & that for ± 3 SD (beyond) = 0.5 + 0.5 = 1%.

Z test gives probability of location of any observation in relation to mean in a normal frequency distribution.

Probability of sample values or means & proportion differing by chance can be determined by shape of sampling.

(5) Probability calculated from table – The calculations of probabilities were also calculated from the various tables like 't' table, chi - square tables etc.

Some other examples of tables –

Life table – survival & dying rate or values

Random table etc.

IMPORTANCE OF PROBABILITY

1. One would like to know the probability of occurrence of sample values by chance so that sample results can be compared with those populations.
2. For this purpose probability of occurrence of biological happening in universe must be understood.
3. Inferences or conclusion drawn after various statistical analyses are based on these probabilities.
4. Uncertainty can be numerically expressed with the help of probability.
5. Fundamental basis of many statistical theories, various distributions are based on probability.

TESTS OF SIGNIFICANCE

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as ***tests of significance.***

As we know the Mean, Standard deviation, proportions etc are very from sample to sample & sample to population (universe). Hence after calculating Mean, Standard deviation, normal distributions etc. it is necessary to draw the conclusion or statistical interference or we can say interpretation of results.

So to get proper conclusion, the significance of difference between the results of our experiment & another worker's experiment or population must be known. The difference observed is expressed in terms of significance or probability or relative frequency of its occurrence by chance & is stated on the basis of sampling distribution.

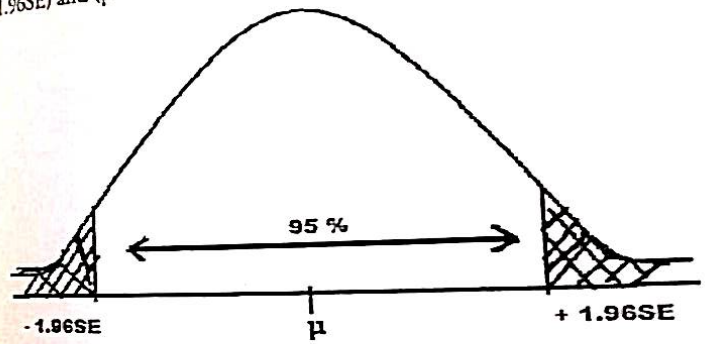
There are 2 methods of drawing the conclusion or knowing the significance of results obtained.

1. Estimation of a population parameter from a sample statistics.
2. The testing of hypothesis about the population parameter.

1. Estimation of a Population Parameter from Sample Statistics:- We cannot cover the entire population & find population parameter (μ). So we calculate it from a sample statistics such as $\bar{X}$.

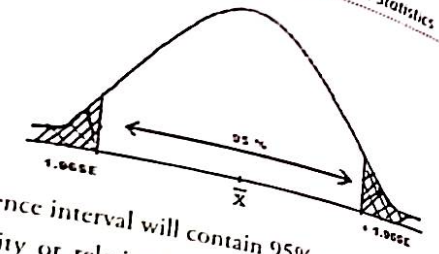
Means $\bar{X}$ (sample statistics) of samples of size 30 or more are normally distributed around the population mean (μ). As they are normally distributed, they must be having certain limits on either side. So, we set up certain limits on both sides of the population mean (μ). These limits are known as "confidence limits" & the range between the two is called as "confidence interval."

By considering the normal distribution, we can say with confidence that 95% of the sample mean ($\bar{X}$) will lie within the confidence limits of (μ – 1.96SE) and (μ + 1.96SE).



(95% confidence interval will contain 95% sample means)

Conversely, the population mean (μ) will also fall within these confidence limits ($\bar{X}$ – 1.96SE) and ($\bar{X}$ + 1.96SE).

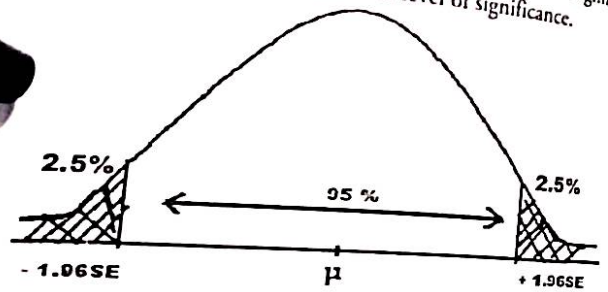


(95% confidence interval will contain 95% population means)

The probability or relative frequency of value lying outside the range mean ±1.96 SE will be 5% or 0.05 out of one.(one in 20times).Same way in mean ±3 SE = 99% confidence limit i.e. only 1% will lie outside.

The values which lie outside the range mean ±1.96 SE or ±2.58 SE have less chance to occur. These values come under the one or on both tails of the normal curve. This region is called as "Level of significance".

Ex– if chance limit is set at Mean ±1.96 SE, the level of significance is 5% or 0.05. it is also called as Critical level of significance.



Similarly if mean is set at mean ±2.58 SE, its level of significance is said to be 1%. A value lying at this limit is called as highly significant & the value lying beyond this area is said to be significantly different.

In most at the statistics, levels of significance are set at 5%, 1% and 0.05%.

Significant or insignificant indicates whether a value is like to occur by chance or it is not occurred by chance.

2. Testing of hypothesis about the population parameter:– *We have seen that, in case if we don't know anything about population parameter.*

we use confidence interval for parameter on the basis of sample from population.

Now consider a different situation. Suppose we have some idea (based on domain knowledge or claim) about what the population parameter will be. Our goal is to test our belief on the basis of collected sample. This procedure is known as "testing of statistical hypothesis."

HYPOTHESIS

It is simply a statement about the population or population parameter.

(1) **Null Hypothesis** – Hypothesis which is normally hypothesized of no difference is called as "Null Hypothesis." Or "Hypothesis of no difference"

According to R.A. fisher – It's a hypothesis which is tested for possible rejection under assumption that, it is true.

This hypothesis nullifies the claim that the experimental result is different from or better that the one observed already.

There is no difference between the results of or statistics of sample & parameter of population or between statistics of two samples.

And it is denoted by H_o.

Ex– The effectiveness of drug A = effectiveness of drug B

Such type of hypothesis indicates – no difference or equality – so called Null hypothesis.

(2) **Alternate Hypothesis** – Any Hypothesis which is complementary to null hypothesis is called as an "Alternate Hypothesis."

This hypothesis states that the sample result is different i.e. greater or smaller than the hypothetical value of population.

And it is denoted by H_1.

Ex– The effectiveness of drug A ≠ effectiveness of drug B

Or

Drug 'B' is more superior or effective than drug 'A'.

In short:-

If we want to test a **Null Hypothesis (H_o)** – that the population has its specified mean 'μ_0'. i.e. $H_o: \mu = \mu_0$

Then the **Alternative Hypothesis (H_1)** is

$H_1 : \mu \neq \mu_0$ (Two tailed)

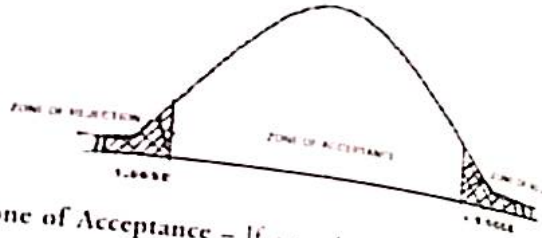
$H_1 : \mu > \mu_0$ (Right tailed)

$H_1 : \mu < \mu_0$ (Left tailed)

Zone of Acceptance and Rejection – By applying the different tests of significance, we either accept H_0 or reject H_0 and vice versa.

To make a minimum error in rejection or acceptance of H_0, we divide sampling distribution or the area under normal curve into two zones i.e.-

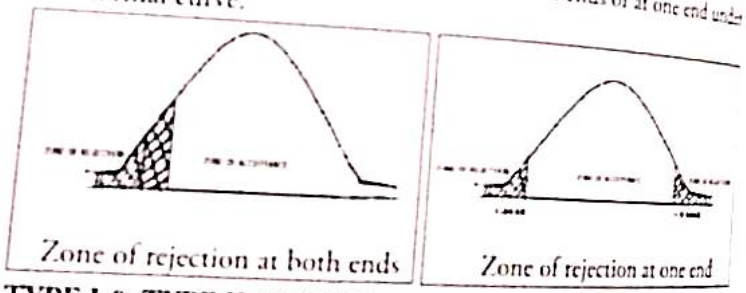
1. Zone of Acceptance
2. Zone of Rejection.



(1) Zone of Acceptance – If a result of sample falls in plain area i.e. within mean ± 1.96 SE, then null hypothesis (H_0) is accepted. Hence this area is called as "Zone of Acceptance for H_0".

(2) Zone of Rejection – If a result of sample falls in shaded area i.e. beyond mean ± 1.96 SE, then H_0 is rejected and H_1 is accepted. Hence this shaded area is called as "Zone of Rejection for H_0".

This Zone of Rejection may distribute at both ends or at one end under the normal curve.



Zone of rejection at both ends

Zone of rejection at one end

TYPE I & TYPE II ERRORS

Generally if result of a sample falls in the zone of acceptance, then null hypothesis is accepted & if result of sample falls beyond it i.e. in zone of rejection then null hypothesis is rejected and alternate hypothesis is accepted.

If we consider mean ±1.96 SE = zone of acceptance [≈95%], then 5% will be the zone of rejection.

But sometimes we do not follow the above rule & accept or reject the null hypothesis even though the above said condition is not present then it leads to error.

So we can say that Hypothesis testing is a decision making procedure. We end up either accepting or rejecting the null hypothesis. There are two types of error which can be happen in this situation.

Type-I Error

- In certain cases, the null hypothesis is rejected even when the estimate falls in the zone of acceptance at 5% level of significance.
- It means we are changing the level of significance from 5% to 6, 8 or 10%.
- It is called as type-I error.
- In short it is a Probability of rejection of H_0, when it is true.
- Type I error is denoted by α.
- It is consider as a more serious error in medical studies.

Type-II Error

The null hypothesis is accepted even if the estimate falls in the zone of rejection at 5% level of significance.

Here we are changing the level of significance from 5% to 4, 3, 2, or 1%.

This is type-II error and denoted by β.

In Short: - If H_0 – True & Accepted = then Result falls within Zone of Acceptance at 5% Level of significance.

If H_0 – False & Rejected = then Result falls in shaded area of rejection at 5% Level of significance.

If H_0 – True & Rejected = then it is TYPE – I error.

If H_0 – False & Accepted = then it is TYPE – II error.

	H_0 True	H_0 False
Accept H_0	Confidence coefficient	Type-II error
Reject H_0	Type-I error	Power

It is clear that we would like both these errors as least as possible. But if we try to minimize one error, other error increases. So we try to keep type-I error (which is serious kind of error) fixed to some low value (0.05 or 0.01) & try to minimize type-II error.

POWER OF TEST: - $(1 - \beta)$

This is probability of rejection null hypothesis when it is false.

It is denoted by $(1 - \beta)$.

TEST STATISTICS

This is function of sample values used to decide whether to reject or accept the null hypothesis. Example – z, t, χ^2

CRITICAL REGION OR REJECTION REGION

The set of all values for which we reject the null hypothesis is known as critical region & it is denoted by C.

ONE TAILED & TWO TAILED TEST

If the critical region is to one side of hypothesized parameter, then test is called "one tailed test" & if it lies to both sides of hypothesized parameter, then test is called as "two tailed test."

Test of Significance – The mathematical procedure used to take decision whether to accept or reject H_0 on basis of computed value of test statistic is called as "Test of significance."

Ex –

Z test, t – test, χ^2 testetc.

Steps to perform test of significance –

1. Set up H_0 and H_1.
2. Decide the Level of significance $\propto$.
3. Calculate approximate test statistic T assuring H_0 is true.
4. If test is one tailed, then find the table value of test statistics at Level of significance $\propto$.
   (1) If $|T_{cal}| > T_{\propto}$ then reject H_0 at $\propto$% Level of significance.

(2) If $|T_{cal}| < T_{\propto}$ then accept H_0 at $\propto$% Level of significance.

If test is two tailed, then find the table value of test statistic at Level of significance $\propto/2$ i.e.

If $|T_{cal}| > T \propto/2$ then reject H_0 at $\propto/2$ Level of significance.

(5) Alternative to 4th step –

In one tail – compute P value of test statistic.

If P Value $> \propto$, then accept H_0 at $\propto$ % Level of significance.

And In two tail -

If P Value $> \propto/2$, then accept H_0 at $\propto$% Level of significance.

(6) Conclusion –

QUESTIONS

1. define probability with its mathematical formula
2. explain laws of probability
3. explain types of probability
4. explain test of significance in details

• • •

PARAMETRIC & NON-PARAMETRIC TESTS

Data can either be continuous, discrete, binary, or categorical.Continuous, or interval, data have units that can be measuredwith a value anywhere between the lowest and the highest value.An example is platelet count. Discrete, or ordinal, data have arank order, but the scale is not necessarily linear. A pain scalefrom 1 to 10 is a good example; a pain score of 8 is not necessarilytwice as bad as 4. Binary data are simply yes/no data: aliveor dead. Examples of categorical, or nominal, data are color orshape. The data are different, but no rank order exists. The testchosen to analyze the data is based on the type of data collectedand some key properties of that data.

Definitions

If the information about the population is completelyknown by means of its parameters then statistical test iscalled **Parametric test.**

Ex: t- test, f-test, z-test, ANOVA Test

If there is no knowledge about the population orparameters, but still it is required to test the hypothesis of the population. Then it is called Non-**parametric test.**

Ex– Mann-Whitney, Wilcoxon rank sum test, Kruskal-Wallis test test

PARAMETRIC TESTS

Parametric tests are more robust and for the most part requireless data to make a stronger conclusion than. However, to use a parametric test, 3 parameters of the data must be true or are assumed.

1. The data need to be normally distributed, which means all data points must follow a bell shapedcurve without any data skewed above or below the mean.

2. The data also need to have equal variance and have the same standard deviation.
3. The data need to be continuous.

NON-PARAMETRIC TESTS

If the data do not meet the criteria for a parametric test (normally distributed, equal variance, and continuous), it must be analyzed with a non parametric test.

If a non-parametric test is required, more data will be needed to make the same conclusion. For this reason, categorical data are often converted to continuous data before analysis.

DIFFERENCE BETWEEN PARAMETRIC TESTS & NON-PARAMETRIC TESTS

Parametric tests	Non-parametric tests
The information about population is completely known.	No information about the population is available.
Specific assumptions are made regarding the population	No assumptions are made regarding the population.
Null hypothesis is made on parameters of the population distribution.	The null hypothesis is free from parameters of the population distribution.
Test statistic is based on the distribution.	Test statistic is arbitrary i.e. not based on distribution.
Parametric tests are applicable only for variable.	It is applied both variable and attributes.
No parametric test exists for Nominal scale data.	Non parametric test do exist for nominal and ordinal scale data.
Parametric test is powerful, if it exists.	It is not so powerful like parametric test.
Examples:- Z test Student 't'- test ANOVA test	Examples:- Chi square test Wilcoxon rank sum test Mann-Whitney test

• • •

Chapter 7

PARAMETRIC TESTS

CHAPTER OVERVIEW

- Z – Test -
  1. Introduction
  2. Types
  3. Steps
  4. Example
- Student 'T'- Test
  1. Introduction
  2. Criteria
  3. Types –
     - (i) One Sample T – Test
     - (ii) Unpaired T - Test
     - (iii) Paired T - Test
- Anova Test –
  1. Steps
  2. Example
- Variance Ratio Test

Z – TEST

It is parametric test.

It is based on Normal Distribution and discovered by "Karl Gows."

When 'Z' test is applied to sampling variability, the difference observed between sample estimate and that of population is expressed in terms of S.E. instead of S.D.

The score of value of ratio between observed difference and Standard Error is Z.

If Z_{cal} falls within mean ± 1.96 SE i.e. Zone of Acceptance then null hypothesis H_0 is accepted.

The distance from mean at which H_0 is rejected is called Level of significance and it falls in zone of Rejection for H_0 which is denoted by 'P.'

P – Indicates Probability or relative frequency of occurrence of difference by chance.

If Z value increases then P Value Decreases.

1. Pre – requisites to apply Z test

a. Samples are randomly selected.

b. Data must be Quantitative.

c. Variables are Normally Distributed.

d. Sample size i.e. N ≥ 30.

Types

(a) One tailed – If the distribution considered only on one side either less than or more than arithmetic mean is called as one tailed Z test.

(b) Two tailed – When both sides of arithmetic mean are considered, then it is called as two tailed Z test.

Steps –

1. Set null hypothesis H_0 and alternate hypothesis H_1.
   (In this case, H_0 is a statement of equality.)
2. Check ∝ i.e. Level of significance.
3. Calculate the test statistic Z under assumption that H_0 is true.

$$Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

4. Find out table value of Z i.e. Z_{table} at given Level of significance ∝
   i.e. $Z_{\propto}$ – One tailed
   $Z_{\propto/2}$ – two tailed.
5. If absolute calculated value i.e. $|Z_{cal}|$ is greater than Z_{table} then we reject the H_0 at ∝% of Level of significance.

If σ is unknown and n ≥ 30 then it is calculated by using 'S' i.e. standard deviation –

$$S = \sqrt{\frac{\Sigma(Xi - \bar{X})}{n-1}}$$

Example No. 1 : A drug manufactures claims that his new drug reduces more than 10 kg weight on average. The 5 patients are taking drug and their weight Level of significances is noted down. Test manufactures claim at 1% Level of significance. (σ = 2)

Given Level of significances of weight – 5, 12, 8, 13, 9.

Solution -

Given – n = 5 σ = 2

μ = 10 ∝ = 1% = 0.01

then,

$$\bar{X} = \frac{5+12+8+13+9}{5} = 9.4$$

Here as H_0: $\mu = 10$

Then, H_1: $\mu > 10$

$$Z_{cal} = Z_{cal} = Z_{cal} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{9.4-10}{2/\sqrt{5}} = \frac{-0.6}{2/2.33} = -0.67 \text{ ---------- (1)}$$

Now table value of Z is calculated from table at 1% of level of significance as -

$$Z_{table} = Z_{\alpha} = 2.33 \text{ ---------- (2)}$$

By comparing equation (1) and (2),

$$|Z_{cal}| < Z_{table}$$

i.e. $|0.67| < 2.33$

So we accept H_0 at 1% level of significance. And hence, we reject the claim of manufacturer.

Example No. 2 : A random sample of 900 children were found to have mean fat fold thickness at triceps of 3.4mm with an Standard deviation of 2.3mm. Can it reasonably regarded as representative sample of population having mean thickness 3.2mm?

Solution :

Given: - $n = 900$, $\bar{X} = 3.4$

$SD = 2.3$, $\mu = 3.2$

Here as,

H_0: $\mu = \mu_0$, here $\mu = 3.2$

H_1: $\mu \neq 3.2$

Then,

$$Z_{cal} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$Z_{cal} = \frac{3.4-3.2}{2.3/\sqrt{900}}$$

$$Z_{cal} = \frac{0.2}{2.3/30}$$

$$Z_{cal} = \frac{0.2 \times 30}{2.3}$$

$$Z_{cal} = \frac{6}{2.3}$$

$$\therefore Z_{cal} = 2.608 \text{ ---------- (1)}$$

Now the table value of Z is calculated from table at 5% of level of significance as -

$$Z_{table} = Z_{\alpha} = 1.96 \text{ ---------- (2)}$$

From equation (1) & (2)

$$|Z_{cal}| > Z_{table}$$

i.e. $|2.608| > 1.96$

So we reject H_0 at 5% level of significance.

Result – it reasonably regarded as representative sample of population is not having mean thickness 3.2mm.

Example No. 3 : 10 individuals are chosen at random from normal population & their weights are found to be 68, 63, 66, 69, 63, 67, 70, 70, 71, and 71. Does this sample adequately represent the population with mean weight less than 66kg? ($\sigma = 5$)

Solution:-

Given: - $n = 10$, $\sigma = 5$, $\mu = 66$

Here as,

H_0: $\mu = 66$

H_1: $\mu \neq > 66$

As we know the formula $Z_{cal} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

The $\bar{X}$ is not given so we have to calculate it first,

$$\bar{X} = \frac{68+63+66+69+63+67+70+70+71+71}{10} = \frac{678}{10} = 67.8$$

Now,

$$Z_{cal} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$Z_{cal} = \frac{67.8-66}{5/\sqrt{10}}$$

$$Z_{cal} = \frac{1.8}{5/3.16}$$

$$Z_{cal} = \frac{1.8 \times 3.16}{5}$$

$$Z_{cal} = \frac{5.69}{5}$$

$$Z_{cal} = 1.138 \text{ ---------- (1)}$$

Now the table value of Z is calculated from table at 5% of level of significance as -

$$Z_{table} = 1.645 \quad \text{---------- (2)}$$

From equation (1) & (2)

$$|Z_{cal}| < Z_{table}$$

$$|1.138| < 1.645$$

So we accept H_0 at 5% level of significance.

Result – Yes, this sample is adequately representing the population with mean 66 kg.

Example No. 4 : a mean relapse time for a certain surgery is claimed to be 8 days. Five patients were observed for relapse. Test the claim at 5% level of significance. (σ = 1 day)

Relapse time: - 4, 9, 7, 12, 15.

Solution :

Given: - $n = 5$, $\sigma = 1$, $\mu = 8$

Here as,

H_0: $\mu = 8$

H_1: $\mu \neq 8$

As we know the formula $Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

The $\bar{X}$ is not given so we have to calculate it first.

$$\bar{X} = \frac{4+9+7+12+15}{5} = \frac{47}{5} = 9.4$$

Now,

$$Z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$Z_{cal} = \frac{9.4 - 8}{1/\sqrt{5}}$$

$$Z_{cal} = \frac{1.4}{1/2.23}$$

$$Z_{cal} = \frac{1.4 * 2.23}{1}$$

$$Z_{cal} = \frac{3.122}{1}$$

$$Z_{cal} = 3.122 \quad \text{---------- (1)}$$

Now the table value of Z is calculated from table at 5% of level of significance as -

$$Z_{table} = Z\alpha/_2\ 1.96 \quad \text{---------- (2) (as it is two tailed so } \alpha/_2)$$

From equation (1) & (2)

$$|Z_{cal}| > Z_{table}$$

$$|3.122| < 1.96$$

So we reject H_0 at 5% level of significance & H_1 is accepted.

STUDENT t - TEST

As Z is based on large sample, but when the sample size is small (i.e. less than 30) then it does not follow Normal Distribution. So Z value will not give correct level of significance for such data.

Hence, there was a need of test of significance for smaller samples. So this test is discovered.

An initial work was done by "W.S. Gossett" Ireland, who published his Research article on this test in penname of "Student test". Therefore this test is known as "Student's t-test or t-test or students test"

The Ratio of observed different between two means of small sample to SE of different in same is denoted by t.

Fishers table gives the highest obtained values of t' under different Probability 'P' in decimal tractions such as (0.01, 0.05, 0.1, 0.001) corresponding to degree of freedom (d*f*)

Probability converted into percentage is stated as level of significance.

P = 0.05 interprets as significant at 5% level. (≅≅ To 95% confidence limits in Z-test)

Probability of t_{cal} is determined by comparing it with value given in row table corresponding to d*f*, derived from number of observations in a sample.

If $t_{cal} > t_{table}$, then Null hypothesis H_0 is rejected & Alternative Hypothesis H_1 is accepted.

If t_{cal} exceeds the value which is given under P = 0.05 in the table, then it is said to be significant at 5% level of significance.

Degrees of freedom: - (d*f*)

The quantity in the denominator which is one less than the independent number of observations in a sample is called as "degree of freedom".

- Degrees of freedom for Unpaired t-test = (n_1+n_2-2)
- Degrees of freedom for Paired t-test = $(n-1)$

Criteria to apply t –test

1. Samples are randomly selected.
2. Data must be Quantitative.
3. Variables are Normally Distributed.
4. Sample size i.e. $n < 30$

Types

1. One sample t-test –
2. Two sample t-test –
   (i) Un-paired t-test
   (ii) Paired t-test.

(A) One sample t – test – If we want to test claim about population mean (μ) on basis of small samples (i.e. $n < 30$) where σ is unknown, then this t-test for one sample is used.

It is based on idea that – $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ which follows t – distribution with $(n-1)$degrees of freedom.

Steps –

1. Set null & alternate hypothesis (i.e.H_0 and H_1).
2. Check for given $\propto$ i.e. Level of significance.
3. Calculate test statistic with assumption that H_0 is true

$$t_{cal} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

4. Find table value of $t_{(n-1)}$ at given level of significance ($\propto$)
   For one tailed - $t_{(n-1)}$, $\propto$
   For two tailed - $t_{(n-1)}$, $\propto/_2$
5. If $|t_{cal}| > t_{table}$ then ,
   H_0 is rejected and H_1 is accepted at $\propto$ % Level of significance.
6. Conclusion.

Ex No 1 – The additional hours of sleep is gained by 8 patients in experiment is given below – 0.7, -1.1, 3.4, 0.8, 2.0, 0.1, -0.2, 3.0

So test the effectiveness of experiment at 5% Level of significance.

Solution -

Given $n = 8$

$\propto = 0.05$

$\mu = 0$

Then first we have to calculate mean $\bar{X}$

$$\bar{X} = \frac{0.7-1.1+3.4+0.8+2.0+0.1-0.2+3.0}{8} = \frac{8.7}{8} = 1.09$$

Now we have to calculate SD, so

X_i	$\bar{X}$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
0.7	1.09	-0.39	0.1521
-1.1	1.09	2.19	4.7916
3.4	1.09	2.31	5.3361
0.8	1.09	-0.29	0.0841
2.0	1.09	0.91	0.8281
0.1	1.09	-0.99	0.9801
-0.2	1.09	1.29	1.6641
3.0	1.09	1.91	3.6481
Total = 8.7	Grand total		17.4888

As

$$S = \sqrt{\frac{\Sigma(Xi-\bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{17.4888}{7}}$$

$$= \sqrt{2.49}$$

$$S = 1.58$$

As $H_0: \mu = 0$

$H_1: \mu > 0$

$$\therefore t_{(n-1)} = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{1.09-0}{1.58/\sqrt{8}} = 1.95 \quad \text{---------- (1)}$$

And

$$t_{table}: t_{0.05,7} = 1.895 \quad \text{---------- (2)}$$

From equation (1) and (2)

i.e. $|t_{cal}| > |t_{table}|$

$|1.95| > |1.89|$

$\therefore H_0$ is rejected at 5% of Level of significance and the Claim is accepted

Ex No 2 – The nicotine content of 5 cigarettes of certain brand showed a mean of 21.2mg with standard deviation of 20.05mg. Test the hypothesis that the average nicotine content of this brand does not exceed 19.7mg (Use $\alpha = 0.05$)

Solution -

Given

$n = 5$ $\quad \bar{X} = 21.2$

$\alpha = 0.05$ $\quad S = 20.05$

$\mu = 19.7$

Here As

$H_0: \mu = 19.7$

$H_1: \mu > 19.7$

We know the formula,

$$t_{(n-1)cal} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$\therefore t_{(n-1)} = \frac{21.2 - 19.7}{20.05/\sqrt{5}}$$

$$\therefore t_{(n-1)} = \frac{1.5}{20.05/2.236}$$

$$\therefore t_{(n-1)} = \frac{1.5 \times 2.236}{20.05}$$

$$\therefore t_{(n-1)} = \frac{3.354}{20.05}$$

$$\therefore t_{(n-1)cal} = 0.167 \quad \text{..... (1)}$$

Now, the degrees of freedom = (n – 1) ∵ (one sample t-test)

$Df = n-1 = 5-1 = 4$

$Df = 4$ ..... (2)

For calculation of table value we have df of 4 at 0.05

$$\therefore t_{(n-1)table} = 2.132 = 2.132 \quad \text{..... (3)}$$

From equation (1) & (3)

$t_{(n-1)cal} < t_{(n-1)table}$

i.e. $0.167 < 2.132$

Result - As we accept the H_0 at 5% level of significance & claim is rejected.

(B) Two sample t-test –

(a) Unpaired t-test:– It is adopted when we want to compare results between two different books.

(b) Paired t-test:– It is used when we want the test of significance of a same sample in different occasions & time like before & after or the interventional readings of same sample (i.e. within same group but at different occasions).

UNPAIRED t-TEST

This test is applied when unpaired data of independent observations of two different groups or samples drawn from two populations.

Criteria to apply unpaired t –test

i. The data must be quantitative & independent.

ii. Samples should be drawn independently & randomly.

iii. The population from which sample is drawn should be normally distributed.

iv. Sample size must be less than 30

Steps to calculate unpaired t-test:-

1. State the Null hypothesis H_0 and alternate hypothesis H_1.

   Here $H_0 = H_1$.

   H_0 is assumed that there is no real different between means of two samples, if samples are taken at random and drawn independently from significance of difference.

2. Find the observed difference between means of two samples $(\bar{X}_1 - \bar{X}_2)$

3. Calculate SE of different between two means – $^{S}(\bar{X}_1 - \bar{X}_2)$

4. then calculate the t value –

$$t_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

Here

(i) If σ is known - then $SE = \sigma\sqrt{\frac{1}{n1}+\frac{1}{n2}}$

(ii) If σ is unknown then we use combined variance.

$$SE = \sqrt{\frac{S1^2}{n1}+\frac{S2^2}{n2}}$$

Then determine the degree of freedom (df) i.e.

$$df = (n_1 + n_2 - 2)$$

Compare the calculated value with table value at particular df

If $|t_{cal}| > t_{table}$ then reject the null hypothesis at $\alpha\%$ level of significance.

i.e. $|t_{cal}| > t_{(n_1+n_2-2),\alpha}$

Or

$|t_{cal}| > t_{(n_1+n_2-2),\alpha/2}$

} then reject null hypothesis at $\alpha\%$ level of significance

Ex – Following are the values of birth weight of high socio-economic and low socio-economic groups. Find whether there is no significant difference between two groups.

Solution:-

	Group A	Group B
n	15	10
$\bar{X}$	2.92	2.26
SD	0.27	0.22

Given As

H_0: $\mu_0 = \mu_1$ (No significant difference)

H_1: $\mu_0 = \mu_1$ (there is significant difference.)

As we know

$$\therefore t = \frac{\bar{X}1-\bar{X}2}{s_{(\bar{X}1-\bar{X}2)}}$$

So first we have to calculate $s_{(\bar{X}1-\bar{X}2)}$

$$\therefore s_{(\bar{X}1-\bar{X}2)} = \sqrt{\frac{(SD1)^2}{n1}+\frac{(SD2)^2}{n2}}$$

$$= \sqrt{\frac{(0.27)^2}{15}+\frac{(0.22)^2}{10}}$$

$$\therefore s_{(\bar{X}1-\bar{X}2)} = 0.098$$

Now we have to calculate t -

$$t_{cal} = \frac{2.92-2.26}{0.098}$$

$$t_{cal} = 6.74 \quad \text{---------- (1)}$$

Then we calculate t_{table} value. For that in unpaired t test–

$$df = n_1 + n_2 - 2$$

$$= 15 + 10 - 2$$

$$df = 23$$

So from table the t_{table} value is.

$$t_{23,0.05} = 1.714 \quad \text{---------- (2)}$$

From equation (1) and (2)

$$|t_{cal}| > t_{23,0.05}$$

Hence we reject the H_0 and Accept H_1.

Calculation of SE of different – $^S(\bar{X}_1 - \bar{X}_2)$ or $\sigma(\bar{X}_1 - \bar{X}_2)$

(1) If μ and σ of population is unknown but the samples are large

$$^S(\bar{X}_1 - \bar{X}_2) \text{ or } \sigma(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(\frac{SD1}{\sqrt{n}}\right)^2 + \left(\frac{SD2}{\sqrt{n2}}\right)^2}$$

$$^S(\bar{X}_1 \text{ s } \bar{X}_2) \text{ or } (\bar{X}_1 \text{ s } \bar{X}_2) = \sqrt{\frac{(SD1)^2}{n1} + \frac{(SD2)^2}{n2}} \text{ OR } \sqrt{\frac{S1^2}{n1} + \frac{S2^2}{n2}}$$

This formula is used when - μ and σ of population is unknown but the samples are large. Thus use of variance of two samples is made in this formula.

$$Z = \frac{\bar{X}1 - \bar{X}2}{s_{(\bar{X}1-\bar{X}2)}}$$

Where $^S(\bar{X}_1 - \bar{X}_2)$ is calculate on the basis of S^2 (variance).

(2) If σ is known but μ is unknown then use this formula –

$$SE = (\bar{X}_1 \text{ s } \bar{X}_2)$$

$$SE = \sqrt{\frac{1}{n1}+\frac{1}{n2}}$$

$$\therefore Z = \frac{(\bar{X}1 - \bar{X}2)}{\sigma\sqrt{\frac{1}{n1}+\frac{1}{n2}}}$$

(3) If Population variance is unknown – Then we find the combined variance of both samples –

$$SE = \sqrt{\frac{(SD)^2}{n1} + \frac{(SD)^2}{n2}}$$

$$SE = \sqrt{SD^2\left(\frac{1}{n1} + \frac{1}{n2}\right)}$$

$$SE = S\sqrt{\frac{1}{n1} + \frac{1}{n2}}$$

Application of SE of difference – To Test $(\bar{X}_1 - \bar{X}_2)$ required applying often in medical practice. As this test is based on Normal Distribution the samples are shown sufficiently large and random.

First set the null hypothesis H_0 and then find the value of Z i.e.

$$Z = \frac{\bar{X}1 - \bar{X}2}{S_{(\bar{X}1-\bar{X}2)}}$$

(b) Paired t-test – It is applied to the paired data of same observations from one sample only.

Criteria to apply Paired t –test

(i) The data must be quantitative.

(ii) Samples should be drawn randomly.

(iii) The data is paired & dependent.

(iv) Sample size must be less than 30.

Application or Uses –

1. To study the role of cause or factor, when observations are made before and after it.

   Like – Guggul or garlic's effect on blood cholesterol level.

2. To compare the effect of two drugs given to the same patient sample on two different occasions.

   Like - number of Hours for which sleep induced by two Hypnotic

3. To study the comparative accurasy of two different instruments

4. To compare the results of two different laboratory techniques.

5. To compare the observations made at two different sites in the body.

   Like – To compare the temprature of Mouth and Rectum.

Steps to calculate paired t-test –

1. State the null hypothesis H_0 and alternative hypothesis H_1
   $H_0 : \mu_1 = \mu_2$
2. Find the difference between each set of paired observations i.e. before and after $(X_1 - X_2 = X)$
3. Then calculate the mean of different $(\bar{X})$
4. Calculate S.D. and then S.E. by – $SE = \frac{SD}{\sqrt{n}}$
5. Calculate t value – $= \frac{\bar{X} - \mu}{S/\sqrt{n}}$
   As H_0 is no different between means of two set, ∴it is Zero '0'
6. Find df – i.e. $df = n-1$
7. Refer 't' table and find the Probability of calculated 't' corresponding to n-1 degree of freedom (df)
8. If Probability (P) is more than 0.05 – then the difference observed has No significance &If P is less than 0.05, then the difference observed is significant.
9. If $|t_{cal}| > t_{table}$ then reject the null hypothesis at $\propto$ % level of significance.

   i.e. $|t_{cal}| > t_{(n-1).\propto}$
   Or
   $|t_{cal}| > t_{(n-1).\propto/2}$

   } then reject null hypothesis at $\propto$ % level of significance

Ex No 1– Following are results of systolic BP before and after treatment of a hypotensive drug of 9 individuals. Test their significance.

Before treatment	122	121	120	115	126	130	120	125	128
After treatment	120	118	115	110	122	130	116	124	125

Solution - Null hypothesis H_0 – the drug is not having the hypotensive effect i.e.

$$H_0 : \mu_1 = \mu_2$$

Alternate hypothesis H_1 – the drug is having the hypotensive effect i.e.

$$H_1 : \mu_1 \neq \mu_2.$$

Before	After	$X_1-X_2=X$	$(X_1-\bar{X})$	$(X_1-\bar{X})^2$
122	120	2	-1	1
121	118	3	0	0
120	115	5	2	4
115	110	5	2	4
126	122	4	1	1
130	130	0	0	0
120	116	4	1	1
125	124	1	-2	4
128	125	3	0	0
		X = 27		24

1. firstly calculate the mean

$$\bar{X} = \frac{X}{n} = \frac{27}{9}$$

$$\bar{X} = 3$$

2. then calculate SD by formula

$$SD = \sqrt{\frac{\Sigma(Xi-\bar{X})^2}{n-1}} = \sqrt{\frac{24}{8}} = \sqrt{3}$$

SD = 1.73

3. now calculate t value –

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

but As H_0 is no different between means of two set,

∴ it is Zero '0' i.e. $\mu = 0$

$$t = \frac{3-0}{1.73/\sqrt{9}} = \frac{3}{1.73/3}$$

$$t = 5.20 \qquad \text{---------- (1)}$$

4. calculate table value i.e. t_{table}

$t_{table} = t_{n-1,\alpha} = t_{8,0.05} = 1.860$ --------- (2)

From equation (1) and (2), we can say that -

$|t_{cal}| > t_{table}$ i.e. 5.20 > 1.860

∴ H_0 – is rejected,and H_1 is accepted at 0.05 level of significance.

It means the given drug is having a hypotensive effect.

Ex No 2 – The weight of 10 tuberculosis patients on admission& at the end of 12 months of treatment are given below. Examone whether the gain in weight is statistically significant?

Sr. no	01	02	03	04	05	06	07	08	09	10
Weight of pt on admission	49	41	37	41	42	37	39	38	41	35
Weight of pt at the end of 12 months t/t	52	43	46	52	46	38	42	41	42	38

Solution – Null Hypothesis = the mean gain in weight is not due to the treatment i.e. $H_0: \mu_1=\mu_2$

Alternate Hypothesis = the mean gain in weight is due to the treatment i.e. $H_1: \mu_1 \neq \mu_2$

Given: -n = 10

Calculate = $\bar{X}$ and SD?

Sr. No.	Weight of patient on Admission (A)	Weight of patients after 12 months (B)	Gain in weight (X_1) $X_1 = B - A$	$X_1-\bar{X}$	$(Xi-\bar{X})^2$
1	49	52	52 - 49 = 3	3 – 4 = -1	01
2	41	43	43 – 41 = 2	2 – 4 = -2	04
3	31	46	46 – 31 = 9	9 – 4 = 5	25
4	41	52	52 – 41 = 11	11 – 4 = 7	49
5	42	46	46 – 42 = 4	4 – 4 =0	00
6	37	38	38 – 37 = 1	1 – 4 = -3	09
7	39	42	42 – 39 = 3	3 – 4 = -1	01
8	38	41	41 – 38 = 3	3 – 4 = -1	01
9	41	42	42 – 41 = 1	1 – 4 = -3	09
10	35	38	38 – 35 = 3	3 – 4 = -1	01

$$\therefore \text{Mean } \bar{X} = \frac{\Sigma Xi}{n} = \frac{3+2+9+11+4+1+3+3+1+3}{10} = \frac{40}{10} = 4$$

$$\therefore \text{Mean } \bar{X} = 4$$

As the sample size is less than 30 (i.e. n < 30) & we have to test the significance within the sample, here paired t-test has to be applied.

Here as $H_0: \mu_1=\mu_2$&

$H_1: \mu_1 \neq \mu_2$

$$\therefore t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

but As H_0 is no different between means of two set,

$\therefore$ it is Zero '0' i.e. $\mu = 0$

$$\therefore t = \frac{\bar{X}-0}{S/\sqrt{n}} = \frac{\bar{X}}{S/\sqrt{n}}$$

Now we have to calculate the SD (s)

$$SD = \sqrt{\frac{\Sigma(Xi - \bar{X})^2}{n-1}}$$

$$\therefore SD = \sqrt{\frac{1+4+25+49+0+9+1+1+9+1}{10-1}}$$

$$\therefore SD = \sqrt{\frac{100}{9}}$$

$$\therefore SD = \sqrt{\frac{100}{9}}$$

$$\therefore SD = \sqrt{11.11}$$

$$\therefore \mathbf{SD = 3.33}$$

Calculation of t value –

$$\therefore t_{cal} = \frac{\bar{X}}{S/\sqrt{n}}$$

$$\therefore t_{cal} = \frac{4}{3.33/\sqrt{10}}$$

$$\therefore t_{cal} = \frac{4}{3.33/3.16}$$

$$\therefore t_{cal} = \frac{4 \times 3.16}{3.33}$$

$$\therefore \mathbf{t_{cal} = 3.795}$$

Now calculate table value:-

$$t_{table} = t_{(n-1).\alpha} = t_{9,\ 0.05} = 1.833$$

$$\therefore \mathbf{t_{9,\ 0.05} = 1.833}$$

By comparing both values,

$$|t_{cal}| > t_{table}$$

i.e. $3.795 > 1.833$

$\therefore$ we reject the H_0 at 5% level of significance& we accept the alternate hypothesis H_1.

Result:- The mean gain in weight is due to the given treatment is proved.

F –test & ANOVA test

As we seen previousely the z test, t-test were used for finding the significance of difference between two samples, but when there are more than two samples then the calculations of z test & t-test becomes very long. So to avoid such a confusing results in testing of hypothesis, the homogenisity of means of more than two sample means can be tested by Analysing the Variability of the groups togethher. (ANOVA)

ANOVA Test (Analysis of Variance)

In previous tests, we used to find the difference between two means by analyzing Standard deviation or standard error, But in this test instead of analyzing SD, we analyses the variance of more than two means and hence this test is called as "Analysis of Variance" i.e. ANOVA.

In this test we compare the means of three or more than three groups simultaneously, by analyzing variability within the group and variability between groups.

It is put forwarded by R.A.Fischer.

This test is applied when more than two samples are drawn from corresponding normal populations.

This test is applied to test the "homogeneity" of means of more than two samples.

Suppose we want to know whether the occupation plays any part in development of hypertension. We take BP randomly of 4 groups –

i.e. 10 – Officers

10 –Lab. Technicians

10 – Clerks

10 – Attendants.

If occupation is playing role in development of hypertension then means of each group will differ significantly.

To test whether 4 means differ significantly then F test or ANOVA test is applied.

Steps to calculate ANOVA test:-

1. State null hypothesis H_0 and alternate hypothesis H_1.
2. Calculate the sum of squares-
   i. Between the groups or classes.
   ii. Within classes.

Sum of square within classes = Total sum of square of entire sample – sum of square between classes.

3. ANOVA table constructed as – (K = number of groups.)

Source of variation	Sum of squares	df	Mean sum of squares	F ratio
Between group	$S \times S_1$	K – 1	MSS_1	$\frac{MSS_1}{MSS_2}$
Within group	$S \times S_2$	n – K	MSS_2	

4. Calculate statistic value of F.
5. Compare the table value of 'F' with $df_{(K-1),(n-K)}$ with F_{cal}.
6. (i) If $F_{cal} > F_{table}$ at given $\propto$% level of significance then H_0 is rejected. And H_1 (significant difference between mean) is accepted.
   (ii) If $F_{cal} < F_{table}$ then H_1 is accepted i.e. samples are drawn from same population.

Ex – Systolic blood pressure values (X) of 4 occupations are given. determine if there is significant difference in mean blood pressure of 4 groups in order to assess the role of occupation in causation of Blood pressure.

Officer X_1	Clerk X_2	Lab. Tech. X_3	Attendants X_4
125	120	120	118
130	122	115	120
135	115	115	118
120	110	130	120
115	125	120	120
120	122	125	115
130	120	122	125

Officer X_1	Clerk X_2	Lab. Tech. X_3	Attendants X_4
135	120	115	125
140	126	126	120
135	120	118	115
Total = 1285	1200	1206	1196
$\bar{X}$ = 128.5	120	120.6	119.6

1. $\Sigma X = 1285 + 1200 + 1206 + 1196$

   $\Sigma X = 4887$

2. Sum of squares of all the 40 observations –

   Sum of sq. $= 125^2+130^2+135^2+\ldots\ldots\ldots+120^2+115^2 = 598751$

3. Total sum of squares $= \Sigma X^2 - \frac{(\Sigma X)^2}{n}$

   $= 598751 - \frac{(4887)^2}{40} = 1681.78$

4. Occupation sum of squares (sum of squares between the classes)

   $= \left[\frac{(\Sigma X1)^2}{n1} + \frac{(\Sigma X2)^2}{n2} + \frac{(\Sigma X3)^2}{n3} + \frac{(\Sigma X4)^2}{n4}\right] - \frac{(\Sigma X)^2}{N}$

   $= \left[\frac{(1285)^2}{10} + \frac{(1200)^2}{10} + \frac{(1206)^2}{10} + \frac{(1196)^2}{10}\right] - \frac{(4887)^2}{40}$

   $= \left[\frac{(1285)^2+(1200)^2+(1206)^2+(1196)^2}{10}\right] - \frac{(4887)^2}{40}$

   $= 538.48$

5. Error sum of squares (Sum of squares within the classes)

   = (Total sum of squares – occupation sum of squares)

   = 1681.78 – 538.48

   = 1143.30

ANOVA table

Square of variance	df	Sum of square	Mean sum	F- ratio of square
Between the occupation	4-1=3	538.48	179.49	5.65
Error	39-3=36	1143.30	31.76	
Total	40-1=39	1681.78		

6. Hence the computed F ratio = $\frac{179.49}{31.76}$

$F_{cal} = 5.65$

7. Calculate F_{table} at df 3 (across) and 36 (vertically) at 5% level of significance in F-table = 2.86

   i.e. $F_{table} = 2.86$

8. As $F_{cal} > F_{table}$, mean BP of the 4 types of employees differ significantly.

And on looking at the means for 4 occupations, we find that mean BP of clerks, laboratory technicians and attendants are comparable but mean BP is apparently highest in the officers.

VARIENCE RATIO TEST

R.A.Fisher was a person who invented this test. Therefore it is called as **"F - test."**

It is a comparison of sample varience.

It involves F- distribution.

First we have to calculate $(S_1)^2$ and $(S_2)^2$ i.e. varience of two samples.

$$\therefore F = \frac{S1^2}{S2^2}$$

$(S_1)^2$ is always greater and placed at the numerator.

The F-table i.e. significance of F is found by refering to F-table.

The degrees of freedom i.e. df will be $n_1 - 1$ and $n_2 - 1$ in two samples.

The F-table gives variance ratio values at different levels of significance at $df(n_1 - 1)$ gives horizontally and $(n_2 - 1)$ given vertically. i.e. $\frac{df(n_1 - 1)}{df(n_2 - 1)}$

If calculated F i.e. $F_{cal} > F_{table}$, then test is significant at 5% (P = 0.05).

QUESTIONS

1. Define parametric test with example
2. Define non parametric test with example
3. Write difference between parametric & non parametric test
4. Write in details Z test
5. Write in details T test
6. Write in details ANOVA test
7. Write short note on Variance ratio Test

• • •

Chapter 8

NON-PARAMETRIC TESTS

CHAPTER OVERVIEW

- Chi-Square Test -
  1. Introduction
  2. Application
  3. Example
- Wilcoxon Test -
  1. Wilcoxon Rank Sum Test
  2. Wilcoxon Signed Rank Test
- Mann-Whitney U-Test
  1. Introduction
  2. Method
  3. Example
- Fisher's Exact Test
  1. Introduction
  2. Application
  3. Example
  4. Use
- Mcnamar's Test
  1. Introduction
  2. Method
  3. Example
- Kruskal-Wallis Test
  1. Introduction
  2. Method
  3. Example

CHI-SQUARE TEST

Introduction –

It is a Non parametric test.

It's not based on any assumption or any distribution variable.

It is commonly used in frequency of data in two or more category.

It is developed by "Karl's Pearson" & has got important application in medical statistics.

Chi – square is from a Greek letter 'Chi' (χ) and pronounced as 'Kye'.

Application –

These are the following three common but very important applications in medical statistics as test of:-

1. Proportion
2. Association
3. Goodness of fit.

(1) Test of Proportion – It's an alternative test to find significant difference in two or more than two proportions.

It is very useful test which can be applied to find significance in same type of data with two or more advantages.

(i) To compare value of two binomial samples even if they are small and less than 30.

Ex.– Incidence of diabetes in 20 non obese patients.
Incidence of diabetes in 20 obese patients.

(ii) To compare the frequency of two multinomial samples.

Ex. – Number of diabetic and non-diabetic patients in group of 40 – 50 kg, 60 – 70 kg etc.

(2) Test of Association – To test the association between two events in binomial or multinomial Samples i.e. two events can often studied for their association.

Ex –

Smoking	-	Cancer
Vaccine	-	Immunity
Wight	-	Diabetes
Cholesterol	-	Coronary Heart disease

Here, there are 2 possibilities that either affected or not affected. And the events are independent or dependent on each other.

When, they are affected or they are dependent then only we can say that they are Associated.

χ^2 test can applied to find association between two events when they are in more than two classes or groups.

Ex – to test the association between number of cigarettes and incidence of cigarettes

If we know the association between two sets of event in table it called as "Association table or contingency table."

When events are in two class or two samples, they are called as "fourfold or four cell or 2x2 contingency table."

Contingency Table:-

		B: Present	B: absent	Total
A	Present	a	b	a + b
	Absent	c	d	c + d
	Total	(a + c)	(b + d)	(a + b + c + d) = N

$$\chi^2 = \frac{(ad-bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

Where N = (a + b + c + d)

Yate's correlation – If value is less than 5,

$$X^2 = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

(3) Test of Goodness of fit – This test is applied to determine that the actual numbers are similar to expected or theoretical numbers.

This test determines whether an observed frequency distribution is differing from theoretical distribution is by chance or the sample is drawn from a difference population.

Rule – If $\chi^2_{cal} > \chi^2_{table}$ then H_0 is rejected.

Steps to calculate χ^2 test –

(1) State the null hypothesis H_0 and alternate hypothesis H_1.

(2) Then obtain an expected frequency (E) by formula –

$$^2 = \frac{\Sigma(Oi-Ei)^2}{Ei}$$

Where rt – rows total
ct – column total
Gt– Grand total

(3) compute the value of χ^2 as -

$$^2 = \frac{\Sigma(Oi-Ei)^2}{Ei}$$

(4) Find out the degree of freedom (d*f*) i.e. d*f* = (c – 1) × (r – 1)

(5) Obtain the table value from χ^2 table.

(6) Compare it with calculated value.

(7) If $\chi^2_{cal} > \chi^2_{table}$ then H_0 is rejected.

Ex No 1 - Apply χ^2 test to find the efficacy of drug from the data given in a table below.

Group	Result		Total
	Died	Survive	
Placebo group	10	25	35
Experimental group	5	60	65
Total	15	85	100

1.Calculate Expected Values (E):-

$$E_{1(\text{died s placebo})} = \frac{15 \times 35}{100} = 5.25$$

$$E_{2(\text{survive s placebo})} = \frac{85 \times 35}{100} = 29.75$$

$$E_{3\ (\text{died s experimental})} = \frac{15 \times 65}{100} = 9.75$$

$$E_{4(\text{survive s experimental})} = \frac{85 \times 65}{100} = 55.25$$

2. χ^2 for each cell –

$$^2_1 = \frac{(Oi-Ei)^2}{Ei} = \frac{(10-5.25)^2}{5.25} = 4.29$$

$$^2_2 = \frac{(Oi-Ei)^2}{Ei} = \frac{(25-29.75)^2}{29.75} = 0.75$$

$$^2_3 = \frac{(Oi-Ei)^2}{Ei} = \frac{(5-9.75)^2}{9.75} = 2.31$$

$$^2_4 = \frac{(Oi-Ei)^2}{Ei} = \frac{(60-55.25)^2}{55.25} = 0.40$$

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\left(\frac{(Oi-Ei)^2}{Ei}\right)$
10	5.25	4.75	22.56	4.29
25	29.75	-4.75	22.56	0.75
5	9.75	-4.75	22.56	2.31
60	55.25	4.75	22.56	0.40

3. Now sum of all χ^2-

$$\chi^2 = \Sigma (\chi^2_1 + \chi^2_2 + \chi^2_3 + \chi^2_4)$$

$$\chi^2 = \Sigma (4.29 + 0.75 + 2.31 + 0.40)$$

$$\chi^2_{cal} = 7.77 \quad \text{---------- (1)}$$

4. Calculations of d*f*-

$$df = (c - 1)(r - 1)$$

$$= (2-1)(2-1)$$

$$df = 1$$

5. Calculation of χ^2_{table} Value -

$\therefore \chi^2_{table}$ at 0.05 = 3.841 ---------- (2)

From equation (1) and (2)

$$\chi^2_{cal} > \chi^2_{table}$$

So H_0 is rejected and hence we can say that drug is effective.

Ex No 2 - Test whether the cancer incidence is associated with smoking at 5% level of significance.

Smoking	Cancer		Total
	Yes	No	
Yes	30	20	50
No	19	31	50
Total	49	51	100

Solution :

H_0 incidence of cancer is independent of smoking.

H_1 incidence of cancer is dependent of smoking.

(1) Calculate Expected Values (E):-

$$E_{11\ (\text{cancers smoker})} = \frac{49 \times 50}{100} = 24.5$$

$$E_{12\ (\text{no cancer s smoker})} = \frac{51 \times 50}{100} = 25.5$$

$$E_{21\ (\text{cancer s non-smoker})} = \frac{49 \times 50}{100} = 24.5$$

$$E_{22\ (\text{no cancer s non-smoker})} = \frac{51 \times 50}{100} = 25.5$$

(2) χ^2 for each cell –

$$^2_1 = \frac{(Oi-Ei)^2}{Ei} = \frac{(30-24.5)^2}{24.5} = 1.234$$

$$^2_2 = \frac{(Oi-Ei)^2}{Ei} = \frac{(20-25.5)^2}{25.5} = 1.186$$

$$^2_3 = \frac{(Oi-Ei)^2}{Ei} = \frac{(19-24.5)^2}{24.5} = 1.234$$

$$^2_4 = \frac{(Oi-Ei)^2}{Ei} = \frac{(31-25.5)^2}{25.5} = 1.186$$

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\left(\frac{(O_i - E_i)^2}{E_i}\right)$
30	24.5	5.5	30.25	1.234
20	25.5	-5.5	30.25	1.186
19	24.5	-5.5	30.25	1.234
31	25.5	5.5	30.25	1.186

(3) Now sum of all χ^2-

$$\chi^2_{cal} = \frac{\Sigma(O_i - E_i)^2}{E_i}$$

$$\chi^2_{cal} = \Sigma (\chi^2_1 + \chi^2_2 + \chi^2_3 + \chi^2_4)$$

$$\chi^2 = \Sigma (1.234 + 1.186 + 1.234 + 1.186)$$

$$\therefore \chi^2_{cal} = 4.84 \quad \text{---------- (1)}$$

(4) Calculations of d*f*-

$$df = (c - 1)(r - 1)$$

$$= (2-1)(2-1)$$

$$df = 1$$

(5) Calculation of χ^2_{table} Value -

$$\therefore \chi^2_{table} \text{ at } 0.05, 1 = 3.841 \quad \text{---------- (2)}$$

From equation (1) and (2)

$$\chi^2_{cal} > \chi^2_{table}$$

Hence, we reject H_0 at 5% level of significance.

Result: - Incidence of cancer is dependent on smoking.

Ex No. 3 – It is claimed that risk of heart disease is uniformly distributed over all age group above 35 years. Test the claim at 5% level of significance with the help of data.

Age group	Number of Heart attack
35 – 40	05
40 – 45	10
45 – 50	12
50 – 55	08
55 – 60	05

Solution :

H_0: data fits well to theory.

H_1: data doesn't fit well to theory.

Age group	No. of Heart attack (O_i)	$E_i = \frac{\Sigma O_i}{n}$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
35 – 40	05	8	9	1.125
40 – 45	10	8	4	0.5
45 – 50	12	8	16	2
50 – 55	08	8	0	0
55 – 60	05	8	9	1.125
Total =	40			4.75

(1) Now sum of all χ^2-

$$\chi^2_{cal} = \frac{\Sigma(O_i - E_i)^2}{E_i}$$

$$\chi^2_{cal} = \Sigma(1.125 + 0.5 + 2 + 0 + 1.125) \quad \text{---------- (1)}$$

$$\therefore \chi^2_{cal} = 4.75$$

Calculation of χ^2_{table} Value -

$$\therefore \chi^2_{table} \text{ at } 0.05, 4 = 9.488 \quad \text{---------- (2)}$$

From equation (1) and (2)

$$\chi^2_{cal} < \chi^2_{table}$$

Hence, we accept H_0 at 5% level of significance.

Result: - data fits well to theory.

WILKOXON TEST

It is non-parametric test.

In this signs of the differences between paired observations & relative ranking are done. It determines whether the before & after data on same patient from a sample is obtained is from same (identical) distribution or not.

As this test uses both the rank i.e. sums & signs of paired difference, that's why it is said to be more efficient.

There are two types of Wilcoxon test

(i) Wilcoxon rank sum test

(ii) Wilcoxon signed rank test.

So we can see this test with example.

(1) Wilcoxon Rank Sum Test

This test says that the null hypothesis tested in a sample is symmetrically distributed around a specified center.

The logic behind the Wilcoxon test is quite simple. The data are ranked to produce two rank totals, one for each condition. If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one.

As a result, the rank totals will be quite different and one of the rank totals will be quite small. On the other hand, if the two conditions are similar, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar and quite large.

The Wilcoxon test statistic "W" is simply the smaller of the rank totals. The SMALLER it is (taking into account how many participants you have) then the less likely it is to have occurred by chance.

A table of critical values of W shows you how likely it is to obtain your particular value of W purely by chance. Note that the Wilcoxon test is unusual in this respect: normally, the BIGGER the test statistic, the less likely it is to have occurred by chance.

Calculations:-

1. The scores exactly equal to the center point are excluded.
2. Remaining scores are ranked like smallest deviation has rank 1.
3. The tied scores are assigned as mean rank.
4. The sum for ranks of scores with positive (+) or negative (-) deviations from central point is calculated.
5. S = smaller of these two rank sums.
6. Then value of S is compared to table of all possible distributions of ranks to calculate probability (P).
7. The sum of all ranks (N) = $\frac{N(N+1)}{2}$

Ex – The recovery time (in days) of a particular disease for 12 patients undergoing two treatments A and B are given. Test the claim that the treatment B is better than treatment A at 5% level of significance.

Solution :

Given –

DRUG 'A'	DRUG 'B'
2	1
3	1
5	3
6	4
1	1
4	2

Solution –

H_0: treatments are equally effective i.e. identically distributed

H_1: treatment B is better.

So arrange the data in ascending order by giving rank.

Sr. No	Time	
1	1	$\frac{1+2+3+4}{4} = 2.5$
2	1	
3	1	
4	1	
5	2	$\frac{5+6}{2} = 5.5$
6	2	
7	3	$\frac{7+8}{2} = 7.5$
8	3	
9	4	$\frac{9+10}{2} = 9.5$
10	4	
11	5	11
12	6	12

Now put the values or ranks in given data –

DRUG 'A'	RANK	DRUG 'B'	RANK
2	5.5	1	2.5
3	7.5	1	2.5
5	11	3	7.5
6	12	4	9.5
1	2.5	1	2.5
4	9.5	2	5.5
TOTAL	48	TOTAL	30

$\therefore W_{cal} = \min(30, 48) = 30$

$\therefore W_{cal} = 30$

And $W_{table} = W_{table(0.05,6)} = 50$

As from above

$$|W_{cal}| < W_{table}$$

Hence, we accept H_0 at 5% level of significance.

$\therefore$ Treatments are equally effective.

(2) Wilcoxon Signed Rank Test

It is used for the analysis of difference between paired observations.

This test is alternative for paired t-test.

Calculations

1. Draw the difference between two paired observations.
2. Then rank the difference by ignoring the sign & give highest rank to highest difference.
3. Again put the signs of each rank as + or –.
4. Sum up the + or – ranks to get W^+ or W^-.
5. If two population means are equal – it indicates that sum of ranks should be equal.
6. And if the difference of sum of ranks is more, then we reject the null hypothesis.

Ex – In order to test efficacy of a diet program, weight of 7 patients were noted down before and after treatment. Test the claim that diet is effective in weight reduction at 5% level of significance. (Assume that data doesn't follow normality)

Weight before	65	79	79	80	83	83	50
Weight after	60	70	78	81	82	83	50

Solution –

Before	After	Difference (d)	Sign
65	60	5	1
79	70	9	1
79	78	1	1
80	81	-1	-1
83	82	1	1
83	83	0	-
50	50	0	-

Arrange the data in ascending order and giving rank number.

H_0: treatment for both population have equal median

H_1: Medians are different

Rank	Number	
-1	1	$\frac{1+2+3}{3} = \frac{6}{3} = 2$
1	2	
1	3	
5	4	= 4
9	5	= 5

Difference |d|	Rank (R_1)	Sign	$R_1 \times$ Sign
|-1|	2	-1	-2
|1|	2	+1	+2
|1|	2	+1	+2
|5|	4	+1	+4
|9|	5	+1	+5
Total =	15		13

$\therefore T^+ = 13$ and $T^- = 2$

$\therefore W_{cal} = \min(13, 2)$

$\therefore W_{cal} = 2$

And $W_{table} = 3$

So from above $|W_{cal}| < W_{table}$ i.e. $2 < 3$

$\therefore$ We reject H_0 at 5% level of significance.

Result: - So we can say that the diet is effective on weight reduction.

Mann-Whitney U-Test

Mann-Whitney-Wilcoxon (MWW) or Wilcoxon Rank-Sum Test

Non-parametric tests are basically used in order to overcome the underlying assumption of normality in parametric tests. Quite general assumptions regarding the population are used in these tests.

A case in point is the Mann-Whitney U-test (Also known as the Mann-Whitney-Wilcoxon (MWW) or Wilcoxon Rank-Sum Test). Unlike its parametric counterpart, the t-test for two samples, this test does not assume that the difference between the samples is normally distributed or that the variances of the two populations are equal. Thus when the validity of the assumptions of t-test are questionable, the Mann-Whitney U-Test comes into play and hence has wider applicability.

Method : The Mann-Whitney U-test is used to test whether two independent samples of observations are drawn from the same or identical distributions. An advantage with this test is that the two samples under consideration may not necessarily have the same number of observations.

This test is based on the idea that the particular pattern exhibited when 'm' number of X random variables and 'n' number of Y random variables are arranged together in increasing order of magnitude provides information about the relationship between their parent populations.

The Mann-Whitney test criterion is based on the magnitude of the Y's in relation to the X's, i.e. the position of Y's in the combined ordered sequence. A sample pattern of arrangement where most of the Y's are greater than most of the X's or vice versa would be evidence against random mixing. This would tend to discredit the null hypothesis of identical distribution.

Assumptions : The test has two important assumptions.

First the two samples under consideration are random, and are independent of each other, as are the observations within each sample.

Second the observations are numeric or ordinal (arranged in ranks).

How to Calculate the Mann-Whitney U : In order to calculate the U statistics, the combined set of data is first arranged in ascending order with tied scores receiving a rank equal to the average position of those scores in the ordered sequence.

Let T denote the sum of ranks for the first sample. The Mann-Whitney test statistic is then calculated using

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T$$

Where n_1 and n_2 are the sizes of the first and second samples respectively.

Ex– An example can clarify better. Consider the following samples.

Sample A

Observation	25	25	19	21	22	19	15
Rank	15.5	15.5	9.5	13	14	9.5	3.5

Sample B

Observation	18	14	13	15	17	19	18	20	19
Rank	6.5	2	1	3.5	5	9.5	6.5	12	9.5

Here,

T = 80.5, (sum of ranks of sample A)

$n_1 = 7$,

$n_2 = 9$.

Hence,

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T$$

$$U = (7 \times 9) + \frac{7(7+1)}{2} - 80.5$$

$$U = 63 + \frac{7 \times 8}{2} - 80.5$$

$$U = 63 + \frac{56}{2} - 80.5$$

$$U = 63 + \frac{56}{2} - 80.5$$

$$U = 63 + 28 - 80.5$$

$$U = 91 - 80.5$$

$$\mathbf{U = 10.5}$$

We next compare the value of calculated U with the value given in the Tables of Critical Values for the Mann-Whitney U-test, where the critical values are provided for given n_1 and n_2, and accordingly accept or reject the null hypothesis.

Even though the distribution of U is known, the normal distribution provides a good approximation in case of large samples.

FISHER'S EXACT TEST

Fisher's exact testis a statistical significance test used in the analysis of contingency tables. Although in practice it is employed when sample sizes are small, it is valid for all sample sizes.

It is named after its inventor, Sir R. A. Fisher, and is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests.

The test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification.

We want to know whether these two classifications are associated.The most uses of the Fisher test involve, like this example, a 2 × 2 contingency table. The p-value from the test is computed as if the margins of the table are fixed and will therefore provide guesses with the correct number in each category.

As pointed out by Fisher, this leads under a null hypothesis of independence to a hyper geometric distribution of the numbers in the cells of the table.

With large samples, a chi-squared test can be used in this situation. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution.

The approximation is inadequate when sample sizes are small, or the data are very unequally distributed among the cells of the table. resulting in the cell counts predicted on the null hypothesis (the "expected values") being low. The usual rule of thumb for deciding whether the chi-squared approximation is good enough is that the chi-squared test is not suitable when the expected values in any of the cells of a contingency table are below 5, or below 10 when there is only one degree of freedom (this rule is now known to be overly conservative).

In fact, for small, sparse, or unbalanced data, the exact and asymptotic

p-values can be quite different and may lead to opposite conclusions concerning the hypothesis of interest. In contrast the Fisher test is, as its name states, exact as long as the experimental procedure keeps the row and column totals fixed, and it can therefore be used regardless of the sample characteristics.

It becomes difficult to calculate with large samples or well-balanced tables, but fortunately these are exactly the conditions where the chi-squared test is appropriate.For hand calculations, the test is only feasible in the case of a 2 × 2 contingency table.

The fisher's exact test uses the formula –

$$P = \frac{[(a+b)!(c+d)!(a+c)!(b+d)!]}{a!b!c!d!N!}$$

Where a,b,c,d = individual frequencies of 2 × 2 contingency table

N = total frequency

Application : The most common reason to avoid chi-square test is because we have small expectation values.

Fisher's exact test is a statistical test used to determine if there are non-random associations between two categorical variables.

Use: The test is most commonly applied to 2 x 2 matrices.

Fisher's Exact Test is a test for independence in a 2 X 2 table. It is most useful when the total sample size and the expected values are small. The test holds the marginal totals fixed and computes the hyper geometric probability that n_{11} is at least as large as the observed value.

It is Useful when E (cell counts) < 5.

McNamar's test

In statistics, McNemar's test is a statistical test used on paired nominal data. It is applied to 2 × 2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (that is, whether there is "marginal homogeneity").

It is named after Quinn McNemar, who introduced it in 1947.

The test is applied to a 2 × 2 contingency table, which tabulates the outcomes of two tests on a sample of n subjects, as follows.

	Test 2 positive	Test 2 negative	Row total
Test 1 positive	a	b	a + b

	Test 2 positive	Test 2 negative	Row total
Test 1 negative	c	d	c + d
Column total	a + c	b + d	n

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same,

i.e. (pa + pb) = (pa + pc) & (pc + pd) = (pb + pd.)

Thus the null and alternative hypotheses are

$$H_0: P_b = P_c$$

$$H_1: P_b \neq P_c$$

Here pa, etc., denote the theoretical probability of occurrence in cells with the corresponding label.

The McNemar test statistic is:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Under the null hypothesis, with a sufficiently large number of discordant (cells b and c), χ^2 has a chi-squared distribution with 1 degree of freedom.

If the χ^2 result is significant, this provides sufficient evidence to reject the null hypothesis, in favour of the alternative hypothesis that pb ≠ pc, which would mean that the marginal proportions are significantly different from each other.

Ex – A researcher attempts to determine if a drug has an effect on a particular disease. Counts of individuals are given in the table, with the diagnosis (disease: *present* or *absent*) before treatment given in the rows, and the diagnosis after treatment in the columns.

The test requires the same subjects to be included in the before-and-after measurements (matched pairs).

	After: present	**After:** absent	Row total
Before: present	101	121	222
Before: absent	59	33	92
Column total	160	154	314

Solution :

In this example, the null hypothesis of "marginal homogeneity" would mean there was no effect of the treatment.

From the above data, the McNemar test statistic:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

$$\chi^2 = \frac{(121-59)^2}{121+59}$$

$$\chi^2 = \frac{(62)^2}{180}$$

$$\chi^2 = \frac{3844}{180}$$

$$\chi^2 = 21.35$$

It has the value 21.35, which is extremely unlikely to form the distribution implied by the null hypothesis ($P < 0.001$).

Thus the test provides strong evidence to reject the null hypothesis of no treatment effect.

Kruskal-Wallis Test

The Kruskal–Wallis one-way analysis of variance by ranks (named after William Kruskal and W. Allen Wallis) is a non-parametric method for testing whether samples originate from the same distribution.

It is used for comparing two or more samples that are independent, and that may have different sample sizes, and extends the Mann–Whitney U test to more than two groups.

The parametric equivalent of the Kruskal-Wallis test is the one-way analysis of variance (ANOVA). While rejecting the null hypothesis of the Kruskal-Wallis test, at least one sample stochastically dominates at least one other sample. The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. Dunn's testwould help analyze the specific sample pairs for stochastic dominance.

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. If the researcher can make the more stringent assumptions of an identically shaped and scaled distribution for all groups, except for any difference in medians, then the null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

Ex – It is of interest to explore the effect that different types of instructions have on solving problems. 12 participants are randomly assigned to one of three groups. The first group of participants was given written instructions. The second group of participants was shown the solution to the puzzle but had no written instructions. The third group of participants was given written instructions and also shown the solution to the puzzle.

Each of the three groups was then asked to solve the puzzle. The number of wrong moves/errors was counted in order to calculate the scores. The results are given below:

Condition 1 (instructions only)		Condition 2 (shown solution and no instructions)		Condition 3 (shown solution and instructions)	
Scores	Ranks	Scores	Ranks	Scores	Ranks
18	1.5	26	8	18	1.5
21	3.5	32	10	21	3.5
24	6.5	34	11	23	5
28	9	36	12	24	6.5

Solution : As the scores come from different participants, it follows that the samples are independent and therefore the Kruskal-Wallis Test holds. In this case the scores are ranked together as an overall set of ranks. These overall ranks can be summed up separately for each condition to produce a rank total for each of the three conditions.

These rank totals are:

$$T_1 = 20.5$$
$$T_2 = 41$$
$$T_3 = 16.5$$

The test statistic for a Kruskal-Wallis test is given by:

$$H = \frac{12}{N(N+1)} \times \sum \frac{(T_i)^2}{n_i} - 3(N+1)$$

Where N is the total amount of participants so N = 12 and

n is the amount of participants in each group so $n_1 = n_2 = n_3 = 4$.

Therefore the test statistic, H, is calculated as follows:

$$H = \frac{12}{N(N+1)} \times \sum \frac{(T_i)^2}{n_i} - 3(N+1)$$

$$H = \frac{12}{12(12+1)} \times \left(\frac{(20.5)^2}{4} + \frac{(41)^2}{4} + \frac{(16.5)^2}{4}\right) - 3(12+1)$$

$$H = \frac{12}{12 \times 13} \times \left(\frac{420.25}{4} + \frac{1681}{4} + \frac{272.25}{4}\right) - 3 \times 13$$

$$H = \frac{12}{156} \times (105.06 + 420.25 + 68.06) - 39$$

$$H = 0.077 \times (593.37) - 39$$

$$H = 45.64 - 39$$

$$H = 6.64$$

The H value of 6.64 is compared to the critical value of 5.69 found in Table E in the appendix. This critical value is calculated by selecting the appropriate size of group's value and p value.

In this case there are 4 participants in each group and the appropriate p value is 0.05 as a 5% significance level is required.

Since 6.64 is greater than 5.69, it can be concluded that since the dependent variable was the amount of mistakes made when solving the puzzles, the higher the participant scores the worse they were at solving the puzzle.

The means of the three groups are 22.75, 32 and 21.5 respectively. This highlights that participants allocated to condition 2 did not perform very well and those participants allocated to condition 1 and condition 3 performed better.

In fact participants in condition 1 and condition 3 made around the same amount of mistakes, suggesting that having written instructions is an important factor regardless of whether the participants have been shown the solution to the puzzle or not.

QUESTIONS

1. Define non parametric test with example
2. Write difference between parametric & non parametric test
3. Explain in details Chi Square test
4. Write short notes on-
   a. Wilcoxon test

b. Mann – Whitney U test
c. Fisher's exact test
d. McNamar's test
e. Kruskal- Wallis test

...

Chapter 9

STATASTICAL SOFTWARE

CHAPTER OVERVIEW

- List of Some Statistics Software Projects
- General Statistics Software Packages
- Statistics And The Computer
- Most Used Software For Multidimensional Data Analysis
- Importance & Application Of Computer In Statistics

List of some Statistics Software Projects

Most of these are open-source software projects

- **Bio Conductor** - software for bioinformatics
- **ESS** - Emacs Speaks Statistics
- **GGobi Data Visualization System**
- **The Omega Project for Statistical Computing**
- **The R Project for Statistical Computing**
- **ROOT** - an object-oriented data analysis framework
- **Statistics Online Computational Resource (SOCR)**
- **StatLib** - a system for distributing statistical software, data sets, and information
- **ViSta**- The Visual Statistics System
- **Weka Machine Learning Project**

General Statistics Software Packages

- **CoStat and CoPlot** - from CoHort Software
- DataDesk
- **Gauss** - from Aptech Systems
- **GLIM** - Genstat from the Numerical Algorithms Group (NAG)
- **JMP** - from SAS
- MedCalc

- Minitab
- **NCSSStatistical Software**
- The R Project for Statistical Computing
- SAS System
- **S-PLUS** - from Insightful Corp. (formerly MathSoft, StatSci)
- SPSS
- Stata
- **STATISTICA** - from StatSoft
- STATGRAPHICS Centurion XVI
- Statistical Solutions (nQuery, BMDP, NCSS)
- Statistix
- Systat
- UNISTAT Statistical Package
- **WINKS** - (formerly KWIKSTAT) from TexaSoft
- XploRe

STATISTICS AND THE COMPUTER

1. Statistics and the computer

There are two different ways in which the computer is changing the field of statistics. First, computers can help us to do what we did before the advent of the computer but in a more efficient way. Second, computers can help us to do things that nobody thought of before the advent of the computer.

To the first category belong statistical data analysis by numerical and graphicalmethods, and simulation; to the second belongs, for example, different computer intensive methods. Another way to categorize the relation statistics-computer is to list the different ways the computer can be used in statistics.

The following are examples of such uses: numerical and graphical data analysis; symbolic computations; simulations; storing statistical knowledge; presentation of results etc.

The close relationship between statistics and computing implies that when one changes the other will also change. The following are some new practical procedures in computing which have turned out to have a great importance for statistics:

(i) The change from mainframe batch computing to personal computing.

(ii) The introduction of multiple dynamic displays.

(iii) The possibility of direct manipulation of graphical objects.

Some trends in statistics are also obviously very much influenced by what has happened in computing.

Examples of such trends are:

(i) Emphasis on exploratory data analysis instead of hypothesis testing;

(ii) The use of computer-intensive methods;

(iii) The introduction of new diagnostic methods.

2. Teaching of statistics and the computer

It is, of course, necessary to consider the use of computers in every course inprobability and statistics, both at school and university levels. However, theintroduction of computers in the teaching process has, in many places, been rather slow.

One obvious reason for this is the considerable cost of establishing an effective computer environment to use in the teaching process. Other reasons for the slowdevelopment are the many difficult problems to be solved before the computer can be incorporated in the teaching of statistics.

Some of these are:

(i) What kind of computer should be used (mainframe, personal, pocket)?

(ii) What make of computers should be used?

(iii) What software should be used?

(iv) What programming language should be used?

(v) What teacher education is needed?

There exists a tremendous amount of statistical software. These software packages are usually produced for use in statistical practice rather than for teaching. But very often they can be used for that purpose, especially if they fulfill the following criteria:

(i) areeasy touse;

(ii) Have good supporting material;
(iii) Are menu-driven;
(iv) Havecolor;
(v) Have good graphics;
(vi) Are interactive;
(vii) Have relevant contents

Courses at university level should give the students familiarity with some common statistical packages, including the new kind with dynamic graphical facilities (e.g. DATADESK, JMP). One possibility is to have a specific course on statistical computing which includes the use of statistical software, but it might be better to incorporate the use of the computer in all courses in probability and statistics.

3. Software for symbolic calculations

A new important feature of computation is the advent of software for symbolic calculations. These computing packages (e.g. Macsyma, Reduce, SMP, Derive, Maple,

Mathernatica) are intended for mathematical calculations but they can turn out to be extremely useful in probability and statistics as well.

A package like Mathernatica offers a large number of mathematical tools for both numerical and graphicalcomputations and contains a lot of useful mathematical knowledge. It also contains some statistical packages for statistical data analysis, with information on statistical distributions. The number of such packages will probably increase rapidly in the future.

These packages can perform algebraic manipulations, integration, differentiation, finding and inverting Laplace and Fourier transforms, linear algebra operations and so on, all of which are important in applied probability and statistics.

Most used software for Multidimensional Data Analysis

(1) Spad
(2) Xl-stat
(3) R
(4) Pspp

(5) Microsoft EXCEL
(6) Open Stat
(7) The Statistics Calculator
(8) Graph Pad –
(a) Graph Pad Prism
(b) In Stat
(c) Stat Mate

(9) SPSS

(1) Spad – Coheris SPAD, a data mining and predictive analytics software, uses company data to anticipate risks, identify opportunities and optimize the different business operations

Advantages

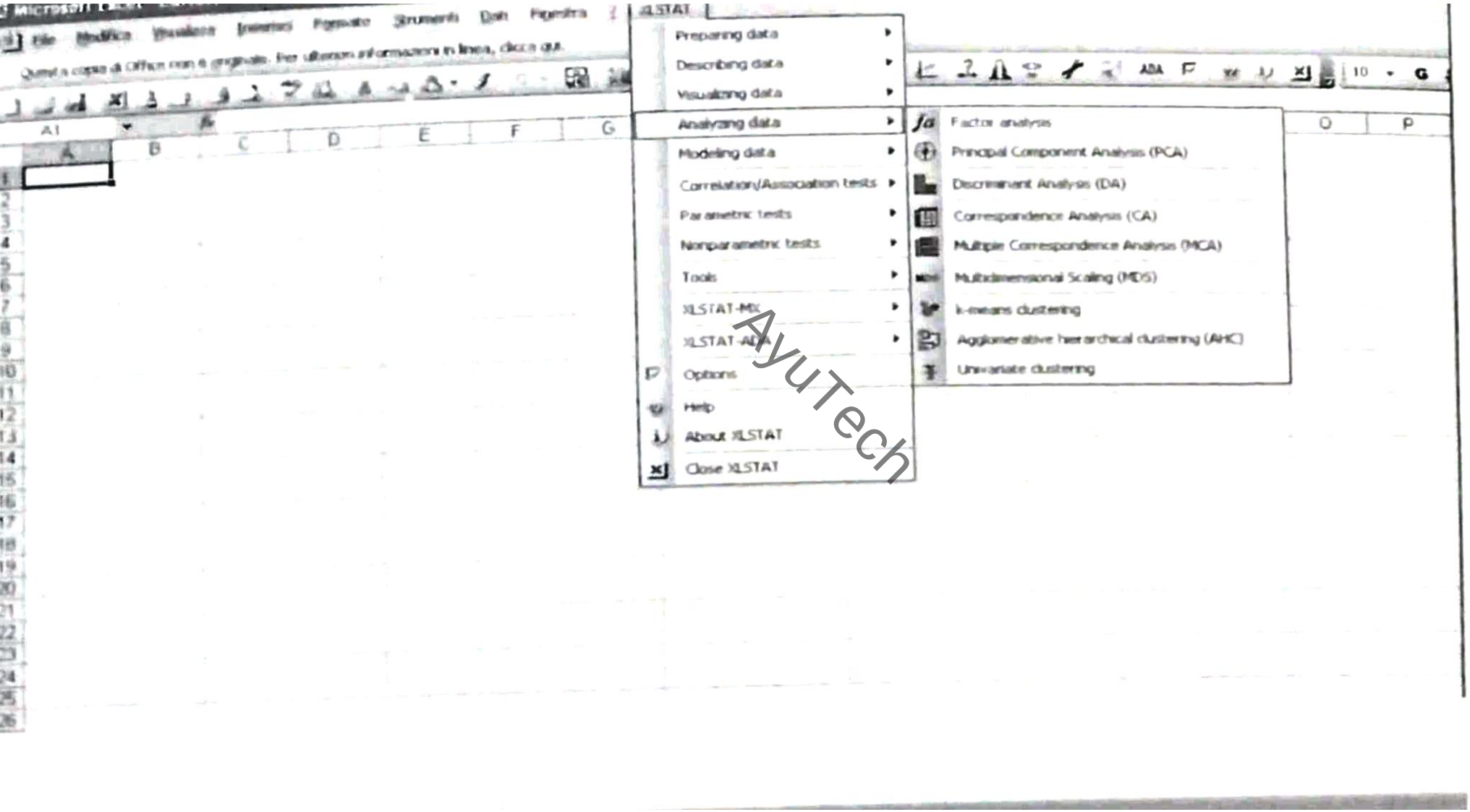
I. It can perform many statistical analyses:
- Descriptive Statistics
- Factorial Analysis
- Classification
- Segmentation
- Textual analysis

II. It has good graphical tools and it is easy to use.

Disadvantages

• Data importation is not direct.

• It is expensive.


XLSTAT
Preparing data
Describing data
Visualizing data
Analyzing data
Modeling data
Correlation/Association tests
Parametric tests
Nonparametric tests
Tools
XLSTAT-MX
XLSTAT-ADA
Options
Help
About XLSTAT
Close XLSTAT
Factor analysis
Principal Component Analysis (PCA)
Discriminant Analysis (DA)
Correspondence Analysis (CA)
Multiple Correspondence Analysis (MCA)
Multidimensional Scaling (MDS)
k-means clustering
Agglomerative hierarchical clustering (AHC)
Univariate clustering
AyuTech

formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

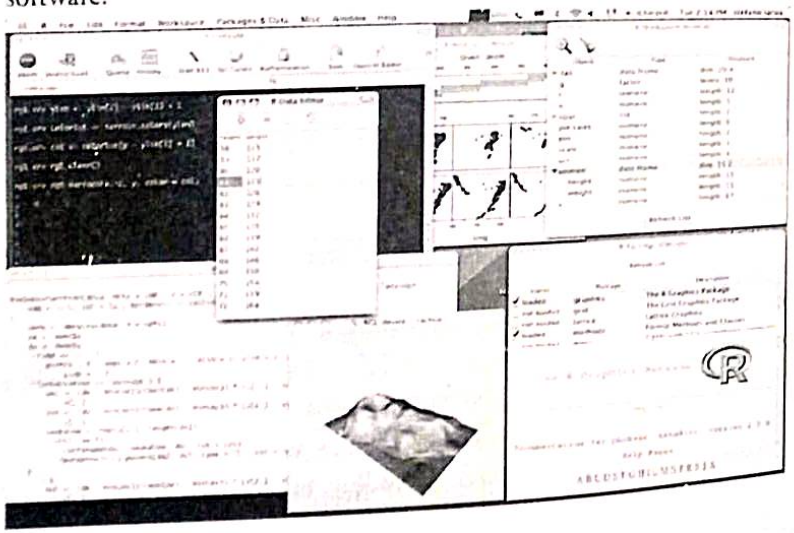
R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including Free BSD and Linux), Windows and MacOS.

The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.



Advantages

Open source.

We can do many statistical procedures.

It's Availability on internet of many routines developed by users.

Disadvantages

Not user-friendly.

(4) PSPP : PSPP is a free software application for analysis of sampled data, intended as a free alternative for IBM SPSS Statistics. It has a graphical user interfaceand conventional command-line interface. It is written in C and uses GNU Scientific Library for its mathematical routines. The name has "no official acronymic expansion".

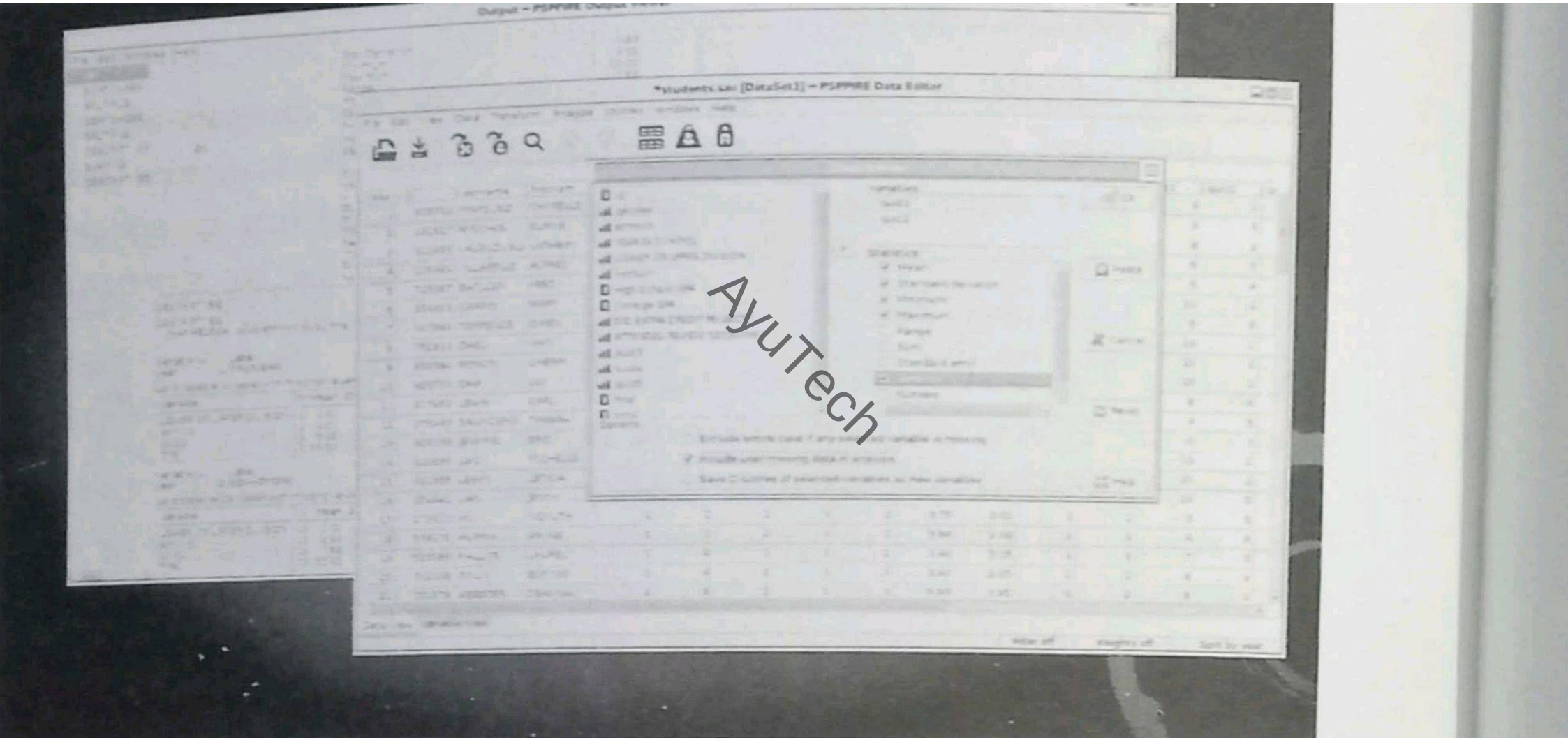
Origins : The PSPP project (originally called "Fiasco") was born at the end of the 1990s as a free software replacement for SPSS, which was a data management and analysis tool, at the time produced by SPSS Inc. The nature of SPSS's proprietary licensing and the presence of digital restrictions management motivated the author to write an alternative which later became functionally identical, but with permission for everyone to copy, modify and share.

Features

This software provides a comprehensive set of capabilities including

- Frequencies,
- Cross-tabs comparison of means (t-tests and one-way ANOVA);
- Linear regression,
- Logistic regression,
- Reliability and re-ordering data,
- Non-parametric tests,
- Factor analysis,
- Cluster analysis,
- Principal components analysis,
- Chi-square analysis and more.

At the user's choice, statistical output and graphics are available in ASCII, PDF, PostScript, SVG or HTML formats. A range of statistical graphs can be produced, such as histograms, pie-charts scree plots and np-charts.


AyuTech

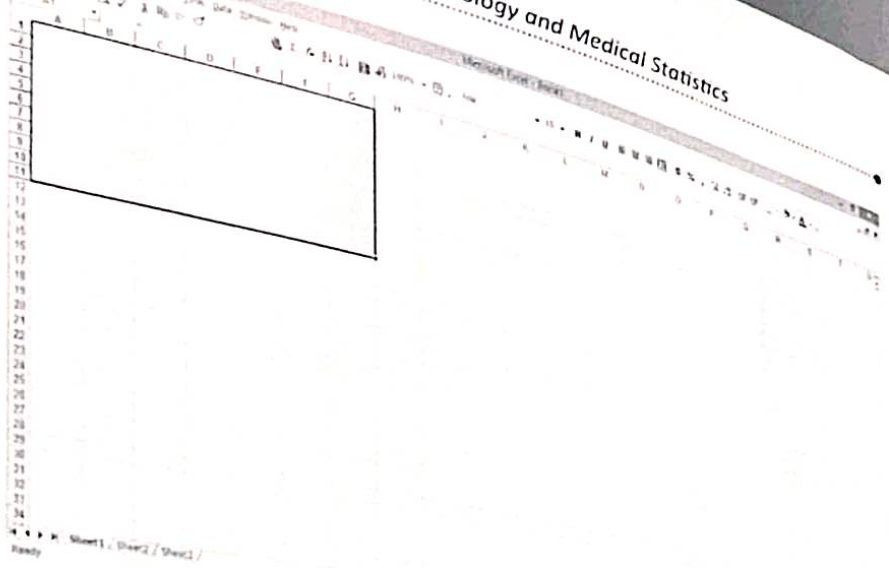
language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

What is Excel?

- Data are organized by worksheets, rows and columns–Worksheet limits are 256 columns and 65,536 total cells
- Cells contain data or formulas with relative or absolute references to other cells
- Direct manipulation of data and flexibility to move data "around" (e.g. sorting, replacing, merging)
- Opens many file types
- Quite useful in prepping files for use in SPSS, SAS or other programs
- Conditional formatting
- Also features macro capabilities, replicating user actions, allowing simple automation of regular tasks.

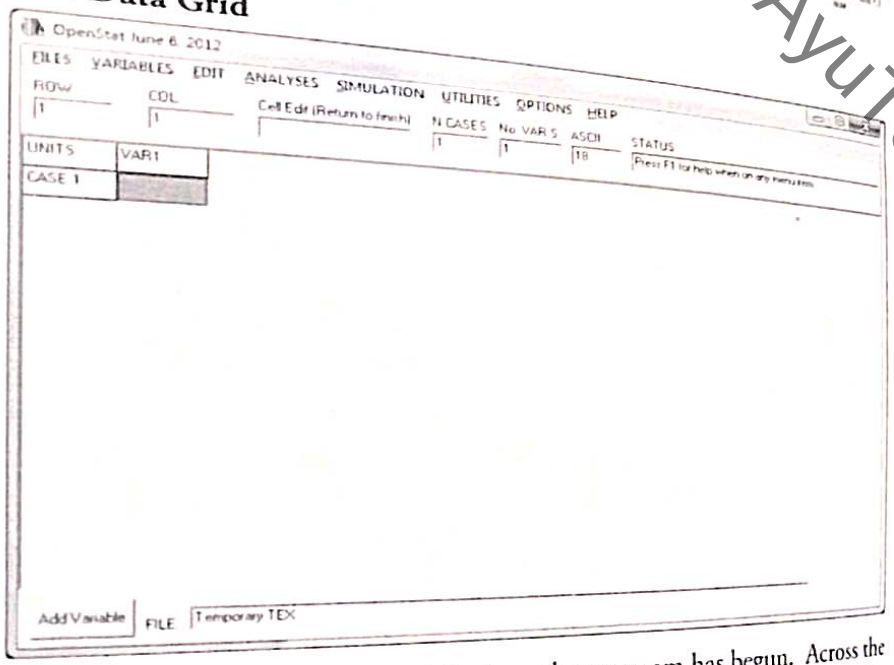
Data Presentation Options in Excel

- Tables and graphs can be exported to a wide variety of software packages
- Can tweak and perfect example graph or table and then replicate by replacing only the data being used
- Main advantage is ability to combine data from multiple sources – not just what is found in the data file
- "Two-for-One" deal - table creation usually puts data into a format that leads to easy graph creation
  - User has control over virtually all aspects of a graph - size, colors, fonts, titles, legends, labels, etc.
  - Can combine graphs with tables and use cell layout to produce more complex presentations
- Final graphs can be of "publication" quality.
- Microsoft Excel has a built-in suite of tools that perform regression analysis and calculate statistical data. These tools are generally installed but are not active. So go to tools in excel sheet & click on



(6) Open Stat Software

The Data Grid



The above shows the data grid when the program has begun. Across the top are the main menus which, when clicked, drop down a list of sub-menus. Below that are some boxes which report the current grid row, column, number of cases, number of variables, the American Standard

Code for Information Interchange (ASCII) for a character entered in a grid cell, and a status indicator. There is also a box for editing the contents of a previously entered cell value. Below the "grid" where data are entered is a button that when pressed adds another variable column (with a default type of floating point values.) There is also a box which indicates the current name of the file. This change after the file has been saved.

The Files Menu : A variety of options exist for saving and opening data files. The preferred method is to use the file extension .TEX which saves not only the data from the grid but also the definition of the variables in the grid. Tab files are useful for importing data from other programs (for example Excel files) or for exporting a file to another program.

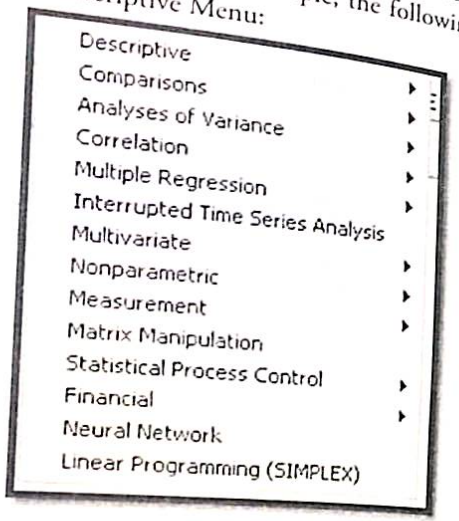
The Variables Menu : The typical user, in creating a new data file, will select the "Define" option. This option lets the user specify the name of a variable (grid column), the type of data in the variable (floating point, integer, string, etc.), the number of decimal fractions, and a value representing a missing value. One can also sort the data in the grid in an ascending or descending order of one of the variables. Occasionally one will need to transform the values of a variable (for example into normally distributed values) or perform a mathematical transformation such as the log of the values. The user can also combine values of several variables to create a new variable or enter an equation to combine and transform multiple variables. If a file contains one variable of values and another variable containing the frequency count of values in the previous variable, one can construct a new file that contains each of the values in one column. Finally, a large file may be split into several files for different analyses or several files may be merged.

The Edit Menu : A typical user will often need to insert a new row or column in the data grid, delete a row or column of the grid, copy and paste row or column data, etc. The Edit menu provides a variety of tools for modifying data in the grid. Excel users, for example, may copy a block of data from an Excel file and paste that block of data into the OpenStat grid. Occasionally a user will need to recode values in one of his or her variables. The recode option provides this capability. In some statistics programs, data which represents group membership may consist of strings such as "Male" or "Female".

OpenStat requires group codes to be integer values. An option exists for creating integer codes from a variable containing string codes. If you are

a user from a country that uses the comma (,) to separate decimal fractions rather than the period (.), you are going to be using the European standard for coding numbers. If you load a file using the USA standard (period separator) you may need to switch the coding to the European standard (comma separator.)

The Analyses Menu : Under the Analyses main menu are listed a number of major sub-menus. Many of the descriptive and analyses procedures produce graphical output. As an example, the following is a plot of X versus Y from the Descriptive Menu:



Clearly, there are too many procedures to describe here. OpenStat contains a large variety of parametric, nonparametric, multivariate, measurement, statistical process control, financial and other procedures. One can also simulate a variety of data for tests, theoretical distributions, multivariate data, etc.

(7) The Statistics Calculator : The Statistics Calculator puts significance testing in the palm of your hand. It's an indispensable tool for report writers who need a quick test to compare means or percent. Get instant answers. No waiting or delays. Make comparisons on the fly and include them in your report. It is an indispensable tool for social science and marketing researchers.

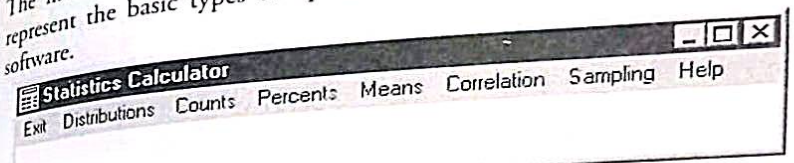
Best of all, the Statistics Calculator is amazingly simple. It installs in seconds and you don't have to be a rocket scientist to use it. In fact, it's

great for beginners. All you do is select a test and fill in a form. There's no hassle and no headaches. Using the software is straightforward and uncomplicated. Plus there's extensive online help to answer all your questions.

Features:-

Main Menu : Statistics Calculator is an easy-to-use program designed to perform a series of basic statistical procedures related to distributions and probabilities. Most of the procedures are called inferential because data from a sample is used to infer to a population.

The menu bar of Statistic Calculator contains eight choices. These represent the basic types of operations that can be performed by the software.



- The **Exit menu** item is used to exit the software.
- The **Distributions menu** item is the electronic equivalent of probability tables. Algorithms are included for the z, t, F, and chi-square distributions. This selection may be used to find probabilities and critical values for the four statistics.
- The **Counts menu** item contains routines to analyze a contingency table of counts, compute Fisher's exact probability for two-by-two tables, use the binomial distribution to predict the probability of a specified outcome, and the Poissondistribution to test the likelihood of observing a specified number of events.
- The **Percent menu** item is used to compare two percent. Algorithms are included to compare proportions drawn from one or two samples. There is also a menu option to calculate confidence intervals around a percent.
- The **Means menu** item is used to calculate a mean and standard deviation of a sample, compare two means to each other, calculate a confidence interval around a mean, compare a sample mean to a population mean, compare two standard deviations to each other, and compare three or more standard deviations.

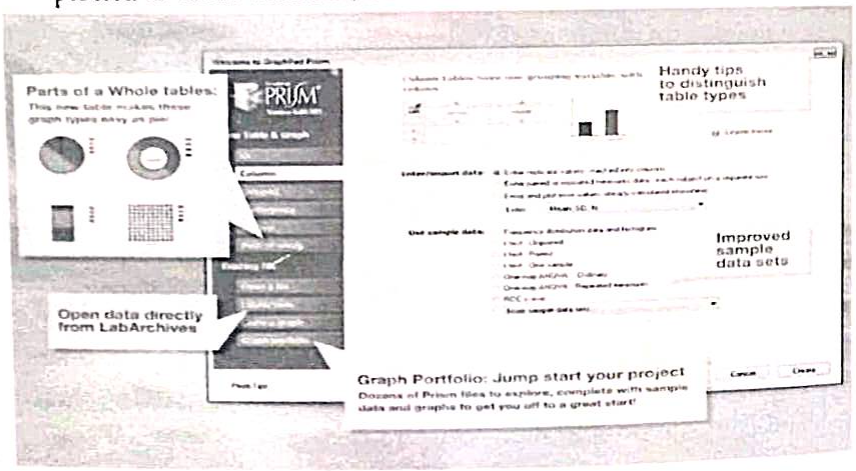
- The **Correlation menu** item is used to calculate correlation and simple linear regression statistics for paired data. Algorithms are included for ordinal and interval data.
- The **Sampling menu** item is used to determine the required sample size for a study. The software can be used for problems involving percent and means.
- The **Help menu** item is used to get this on-line help.

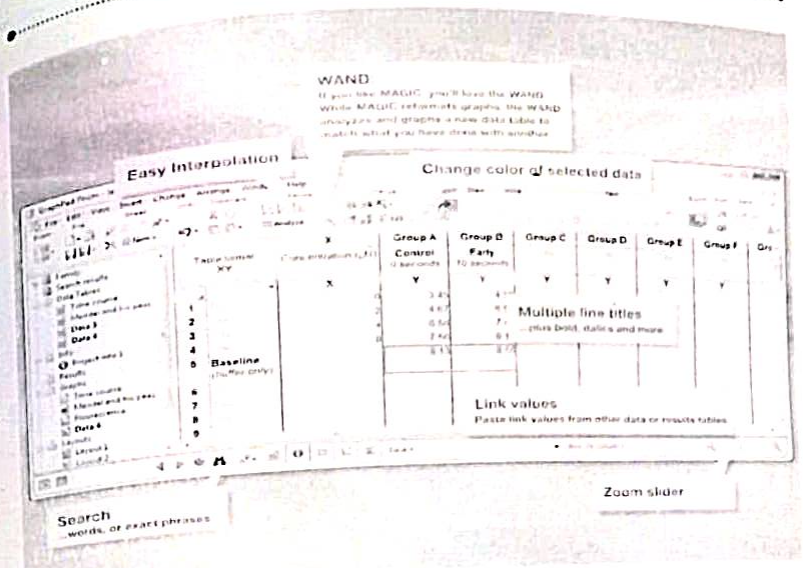
(8) Graph pad

(A) Graphpad Prism : GraphPad Prism is commercial scientific 2D graphing and statistics software published by GraphPad Software, Inc., a privately held California corporation. Prism is available for both Windows and Macintosh computers.

Features

- It provides statistical guidance for novices.
- Analysis checklists review if an appropriate analysis was performed.
- Nonlinear regressions with many options (remove outliers, compare models, compare curves, interpolate standard curves, etc.).
- Live links. When data are edited or replaced, Prism automatically updates the results and graphs.
- Analysis choices can be reviewed, and changed, at any time.
- Automatic error bars. Raw data (replicates) can be entered, and then plotted as mean with SD, SEM or confidence interval.





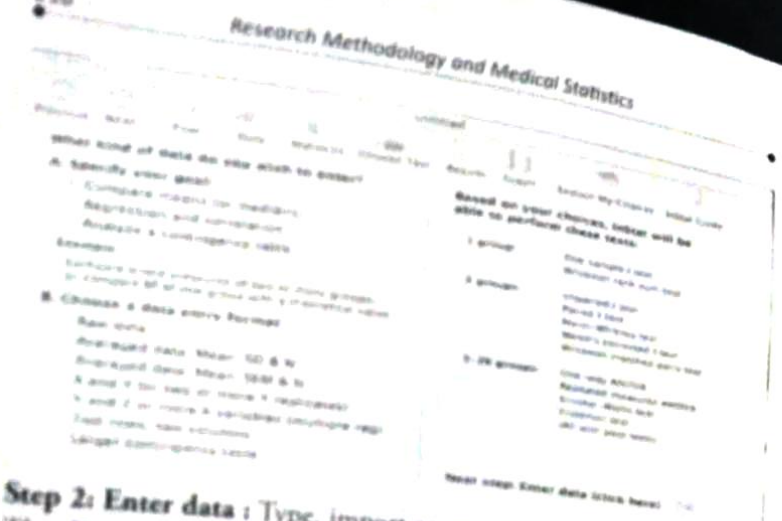
(B) Graph Pad InStat : InStat is designed for statistical novices. It helps you pick an appropriate test by asking questions about the data. It presents results in simple paragraphs, with a minimum of statistical jargon.

InStat's help screens review the use of each test and explain every portion of the results, and provide unique analysis checklists to confirm that the data have not violated any assumptions of the test, and that the test answers the right question.

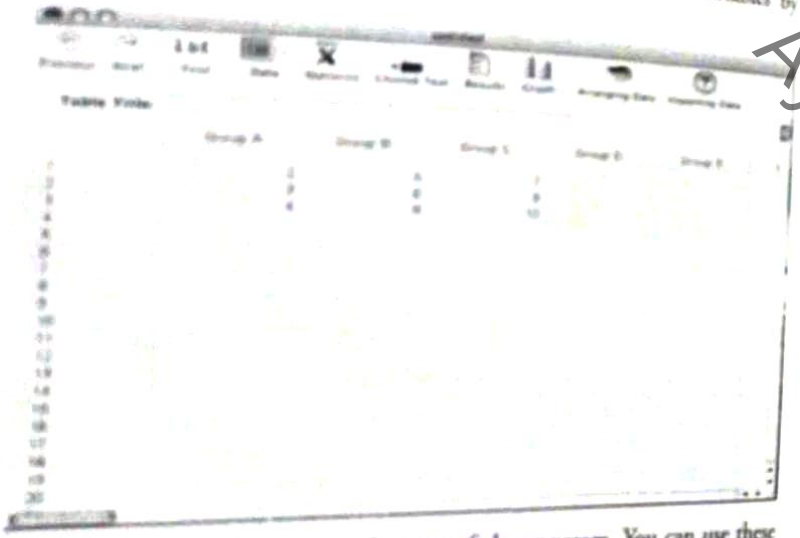
This overview, shown here will demonstrate how easy it is to perform one-way ANOVA using InStat in just a few minutes.

Step 1: Choose data format : Before you can enter data, you first have to tell InStat what kind of data table you need. This important step makes InStat unique. Once you've chosen the right kind of data table for your experiment, InStat will be able to guide you to choose an appropriate statistical test. InStat offers three goals on the top of the screen, with more choices below. Based on your choices, you'll be able to perform different tests as shown on the right. The three goals are distinct, and you shouldn't have a problem telling InStat what kind of data you have.

For this example, we will select Compare means and raw data.

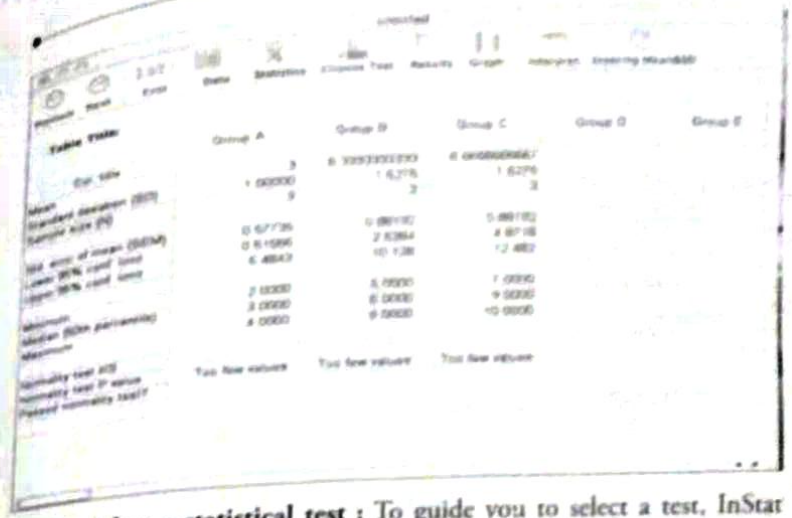


Step 2: Enter data : Type, import or paste data onto InStat's data table. The Data menu lets you transform data or create new variables by combining columns.

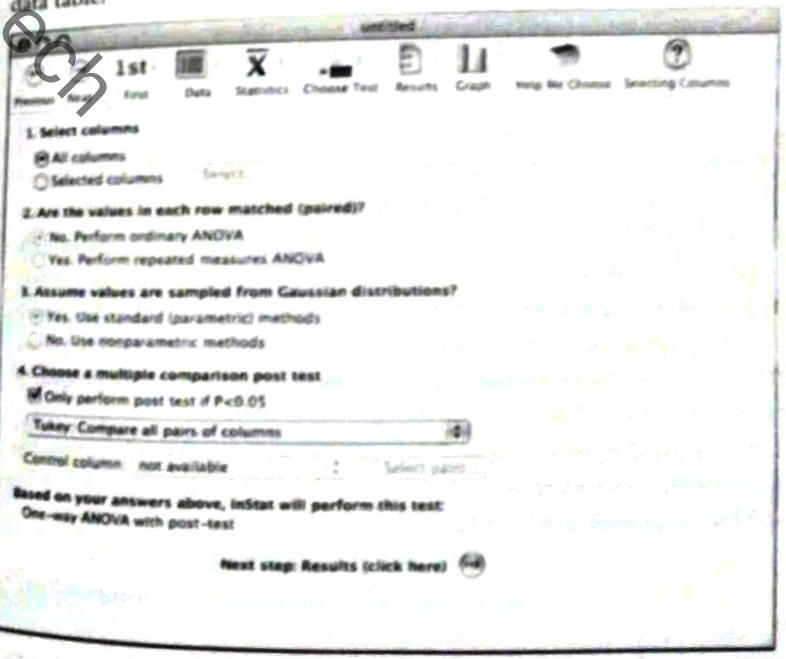


Note the six step buttons at the top of the program. You can use these buttons to move from step to step in any order or you can use the arrow buttons to move in sequence. The third button goes to the next step.

Step 3: Summary statistics : For each column, view the mean, SD, SEM, confidence interval, etc. You can also enter the mean, sample size, and SD (or SEM) here if you have already analyzed data in another program.



Step 4: Select a statistical test : To guide you to select a test, InStat presents four questions. You would see different choices if you entered a different number of columns of data, or if you created a different kind of data table.



One-way Analysis of Variance (ANOVA)

The P value is 0.0067, considered very significant.
Variation among column means is significantly greater than expected by chance.

Tukey-Kramer Multiple Comparisons Test

If the value of q is greater than 4.339 then the P value is less than 0.05.

Comparison	Mean Difference	q		P value
Column A vs Column B	-3.333	4.201	ns	P>0.05
Column A vs Column C	-5.667	7.141	**	P<0.01
Column B vs Column C	-2.333	2.941	ns	P>0.05

Difference	Mean Difference	95% Confidence Interval From	To
Column A - Column B	-3.333	-6.776	0.1096
Column A - Column C	-5.667	-9.110	-2.224
Column B - Column C	-2.333	-5.776	1.110

(9) SPSS

What is SPSS? : Originally it is an acronym of Statistical Package for the Social Science but now it stands for Statistical Product and Service Solutions.

It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.

Advantages

- It is Easy to use, similar to excel.
- It is simple to find manual or tutorial on internet that shows how to use it.
- It has a large diffusion also if it is not open source.

Disadvantages

- For some analyses it has less option than other packages.
- The Trial version is only for 14 days and limited license.

Importance & application of Computer in Statistics

Computer science uses statistics in many ways to guarantee products available on the market are accurate, reliable, and helpful. The following terms and definitions are a brief listing of areas in computer science that use statistics to varying degrees at various times.

1. **Data Mining** is the analysis of information in a database, using tools that look for trends or irregularities in large data sets.
2. **Data Compression** is the coding of data using compact formulas, called algorithms, and utilities to save storage space or transmission time.
3. **Speech Recognition** is the identification of spoken words by a machine. The spoken words are turned into a sequence of numbers and matched against coded dictionaries.
4. **Vision and Image Analyses** use statistics to solve contemporary and practical problems in computer vision, image processing, and artificial intelligence.
5. **Human/Computer Interaction** uses statistics to design, implement, and evaluate new technologies that are useable, useful, and appealing to a broad cross-section of people.

6. **Network/Traffic Modeling** uses statistics to avoid network congestion while fully exploiting the available bandwidth.
7. **Stochastic Optimization** uses chance and probability models to develop the most efficient code for finding the solution to a problem.
8. **Stochastic Algorithms** follow a detailed sequence of actions to perform or accomplish a task in the face of uncertainty.
9. **Artificial Intelligence** is concerned with modeling aspects of human thought on computers.
10. **Machine Learning** is the ability of a machine or system to improve its performance based on previous results.
11. **Capacity Planning** determines what equipment and software will be sufficient while providing the most power for the least cost.
12. **Storage and Retrieval techniques** rely on statistics to ensure computerized data is kept and recovered efficiently and reliably.
13. **Quality Management** uses statistics to analyze the condition of manufactured parts (hardware, software, etc.) using tools and sampling to ensure a minimum level of defects.
14. **Software Engineering** is a systematic approach to the analysis, design, implementation, and maintenance of computer programs.
15. **Performance Evaluation** is the process of examining a system or system component to determine the extent to which specified properties are present.
16. **Hardware Manufacturing** is the creation of the physical material parts of a system, such as the monitor or disk drive.

Statistics is essential to the field of computer science in ensuring effectiveness, efficiency, reliability, and high-quality products for the public.

QUESTIONS

1. Write Importance of software in statistics
2. Write Short notes on SPSS
3. Write generally used software in statistics
4. Write importance & application of computer in statistics.

• • •